

Analysing and testing the ratio of infection rate and its truthfulness



MARIYA TSVETANOVA TODOROVA



Data Analysis and visualisation (ACCB4002)

Lecturer Anupam Mazumdar

University of Wales Trinity Saint David, London

Date: 04/01/2021

Word limit: 2000

Table of Contents

List of Tables: _____	2
List of figures: _____	3
Introduction: _____	4
Preamble: _____	6
What is a t-test and in what does it consist 34words _____	6
Step-by-step development: _____	7
Dependent variables, correlation, and linear regression: _____	9
Standard error. What is and how to predict when it occurs: _____	9
Null Hypothesis and Alternative Hypothesis: _____	10
T-test. What is and what is used for: _____	11
Type of t-test run, motivations and analysis of the results outputted: _____	14
Accepting or rejecting the Null Hypothesis: _____	15
T-test run with significant value of 0.10: _____	15
T-test run with significant value of 0.01: _____	15
T-test run with a significant value of 0.02 _____	16
Null Hypothesis and Alternative Hypothesis regarding this test: _____	17
Visualisation of the age range and frequency of the infection between the age ranges _____	18
Conclusion: _____	19

List of Tables:

Table number:	Description of the table:	Page number:
1	List of figures	3
2	Introduction	4
3	Preamble	6
4	What is a t-test and in what does it consist	6
5	Step-by-step development	7
6	Dependent variables, correlation, and linear regression	9
7	Standard error. What is and how to predict when it occurs	9
8	Null Hypothesis and Alternative Hypothesis	10
9	T-test. What is and what is used for	11
10	Type of t-test run, motivations and analysis of the results outputted	14
11	Accepting or rejecting the null hypothesis	15
12	T-test run with significant value of 0.10	15
13	T-test run with significant value of 0.01	15
14	T-test run with significant value of 0.02	16
15	Null Hypothesis and Alternative Hypothesis regarding this test	17
16	Visualisation of the age range and frequency of the infection between the age ranges	18
17	Conclusion	19

List of figures:

<i>Figure number:</i>	<i>Description of the figure:</i>	<i>Page number:</i>
1	Performance of the t-test step-by-step	5
2	Differences between Inferential and Descriptive Statistics	6
3	Difference between the two types of correlations	7
4	sampling methods	7
5	Difference between Dependent and Independent variables	9
6	Standard error types	9
7	Type of standard error occurring	9
8	formula to estimate the value of the standard error	10
9	Null Hypothesis and Alternative Hypothesis, and how to determine which one to use	10
10	Types of t- test and its relative formulas to be followed	11
11	Distribution table	12
12	reminder for t_{obt} value	13
13	t-test two sample, assuming Unequal variances with p value=0.10	14
14	Significant level value rules. When to accept or reject a H_0	14
15	T-test two sample, assuming Unequal variance with p value=0.01	15
16	T-test two sample, assuming Unequal variance with p value=0.02	16
17	Infographic showing the differences between White and BAME patients' rate of contagion	17
18	Infographic showing the frequency of infection rate divided by age range	18

Data Analysis and Visualisation: Assignment -2

Introduction:

Considering the given data table, and the operation required to be executed I will first analyse and explain the processes that will guide me to execute and select the appropriate t-test for this data set. I will use the support of references to online sources/videos and step-by-step demonstrations of my development of the task in excel, and by explaining with the appropriate statistical expressions each component used to establish if the infection rate is $\geq$; $\leq$; $=$; $\neq$ between the two groups.

Data Analysis and Visualisation (ACCB4002):

patient N	White patient	BAME		x $= X - \bar{x}$	y $= Y - \bar{y}$	$x \cdot y$ $= (x) \cdot (y)$	x^2	y^2
1	20	23		-16.18	-13.74	222.3132	261.7924	188.7876
2	22	34		-14.18	-2.74	38.8532	201.0724	7.5076
3	25	44		-11.18	7.26	-81.1668	124.9924	52.7076
4	26	45		-10.18	8.26	-84.0868	103.6324	68.2276
5	29	29		-7.18	-7.74	55.5732	51.5524	59.9076
6	27	27		-9.18	-9.74	89.4132	84.2724	94.8676
7	48	35		11.82	-1.74	-20.5668	139.7124	3.0276
8	43	45		6.82	8.26	56.3332	46.5124	68.2276
9	55	44		18.82	7.26	136.6332	354.1924	52.7076
10	43	30		6.82	-6.74	-45.9668	46.5124	45.4276
11	27	40		-9.18	3.26	-29.9268	84.2724	10.6276
12	34	50		-2.18	13.26	-28.9068	4.7524	175.8276
13	44	50		7.82	13.26	103.6932	61.1524	175.8276
14	50	50		13.82	13.26	183.2532	190.9924	175.8276
15	37	41		0.82	4.26	3.4932	0.6724	18.1476
16	35	43		-1.18	6.26	-7.3868	1.3924	39.1876
17	33	47		-3.18	10.26	-32.6268	10.1124	105.2676
18	40	32		3.82	-4.74	-18.1068	14.5924	22.4676
19	42	31		5.82	-5.74	-33.4068	33.8724	32.9476
20	48	20		11.82	-16.74	-197.8668	139.7124	280.2276
21	49	26		12.82	-10.74	-137.6868	164.3524	115.3476
22	30	28		-6.18	-8.74	54.0132	38.1924	76.3876
23	33	30		-3.18	-6.74	21.4332	10.1124	45.4276
24	43	39		6.82	2.26	15.4132	46.5124	5.1076
25	50	50		13.82	13.26	183.2532	190.9924	175.8276
26	50	50		13.82	13.26	183.2532	190.9924	175.8276
27	32	33		-4.18	-3.74	15.6332	17.4724	13.9876
28	20	45		-16.18	8.26	-133.6468	261.7924	68.2276
29	30	47		-6.18	10.26	-63.4068	38.1924	105.2676
30	40	20		3.82	-16.74	-63.9468	14.5924	280.2276
31	44	32		7.82	-4.74	-37.0668	61.1524	22.4676
32	35	22		-1.18	-14.74	17.3932	1.3924	217.2676
33	28	28		-8.18	-8.74	71.4932	66.9124	76.3876
34	40	30		3.82	-6.74	-25.7468	14.5924	45.4276
35	43	34		6.82	-2.74	-18.6868	46.5124	7.5076
36	38	32		1.82	-4.74	-8.6268	3.3124	22.4676
37	33	44		-3.18	7.26	-23.0868	10.1124	52.7076
38	34	47		-2.18	10.26	-22.3668	4.7524	105.2676
39	50	30		13.82	-6.74	-93.1468	190.9924	45.4276
40	22	44		-14.18	7.26	-102.9468	201.0724	52.7076
41	37	29		0.82	-7.74	-6.3468	0.6724	59.9076
42	45	40		8.82	3.26	28.7532	77.7924	10.6276
43	20	30		-16.18	-6.74	109.0532	261.7924	45.4276
44	44	40		7.82	3.26	25.4932	61.1524	10.6276
45	33	24		-3.18	-12.74	40.5132	10.1124	162.3076
46	20	50		-16.18	13.26	-214.5468	261.7924	175.8276
47	38	46		1.82	3.26	16.8532	3.3124	85.7476
48	35	30		-1.18	-6.74	7.9532	1.3924	45.4276
49	25	27		-11.18	-9.74	108.8932	124.9924	94.8676
50	40	50		3.82	13.26	50.6532	14.5924	175.8276
N=50	$\Sigma x = 1809$	$\Sigma y = 1837$				$\Sigma xy = 308.34$	$\Sigma x^2 = 4347.38$	$\Sigma y^2 = 4253.62$
$\bar{x} = \frac{\Sigma x}{N}$		$\bar{x} = 1809:50 =$		36.18				
$\bar{y} = \frac{\Sigma y}{N}$		$\bar{y} = 1837:50 =$		36.74				
$cov x y = \frac{\Sigma xy}{N - 1}$		$cov x y = 308.34:49 =$		6.292653061				
$var x = \frac{1}{N - 1} \cdot \Sigma x^2$		$var x = \frac{1}{49} \cdot 4347.38 =$		88.72204082				
$var y = \frac{1}{N - 1} \cdot \Sigma y^2$		$var y = \frac{1}{49} \cdot 4253.62 =$		86.80857143				
$SDx = \sqrt{var x}$		$SDx = \sqrt{88.72204082} =$		9.419237804				
$SDy = \sqrt{var y}$		$SDy = \sqrt{86.80857143} =$		9.317111754				
$r_{xy} = \frac{cov x y}{SDx \cdot SDy}$		$r_{xy} = \frac{6.292653061}{9.419237804 \cdot 9.317111754} =$		0.0717029				
$t_{obt} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{SD_1^2 + SD_2^2}{n}}}$		$t_{obt} = \frac{36.18 - 36.74}{\sqrt{\frac{(9.419237804)^2 + (9.317111754)^2}{50}}} =$		-0.29888				
$df = n_1 + n_2 - 2 =$		$df = 50 + 50 - 2 =$		98				

Figure n° 1: Performance of the t-test step-by-step. (Credits: created by the author).

Preamble:

To make clear and understandable all the concepts of the process of my data analysis we must know some basic rules which will bring us to the application of the t-test. To do so, we need to also understand what a t-test exactly is.

What is a t-test and in what does it consist

It's an inferential statistic [1] tool, used to understand if the mean (μ) of two samples differ to a statistically significant degree. T-test also helps to test hypothesis and answer questions based on its results. 34words

	Inferential Statistics:	Descriptive Statistics:
What is:	Is commonly, a kind of test run based on our predictions.	Is commonly, a set of methods that helps to organize and summarize information.
Consist of:	Methods that allow to reach conclusions and make estimations about populations based on information taken by a sample.	Methods dealing with collecting, tabulation, summarising, and presentations of data.
Gives us:	Methods for drawing and measuring the reliability of conclusion about population based on information obtained from a sample.	Methods for organizing and summarising information; includes the production of graphs, charts, and tables.
Allows us to:	Identify methods that include point estimation, interval estimation and hypothesis testing (t-tests); all based on probability.	Calculate various descriptive measures such as averages, measures of variations and percentiles.
What does it do:	Essentially makes predictions.	Essentially describes data.

Figure n° 2: Differences between Inferential and Descriptive Statistics. (Credits: created by the author).

The term “population” is usually referred to the group of all individuals/items took in consideration in a statistical way. When we discuss about samples or population samples, we refer to a group of it, taken under examination to learn about the whole population.

T-test usually considers samples of population.

To answer our questions about this assignment, we must choose one of the following:

- T-test for paired two sample means
- T-test for two sample assuming equal variances
- T-test for sample assuming unequal variances

To develop our analysis based on the t-test we should now choose how to run the test: following different formulas and analysing the result obtained through performing step-by-step formulas, or by using the automated performing via excel. While developing my assignment I choose to do both to confirm my decision.

For the step-by-step execution, we must start by applying the **Pearson Correlation Coefficient** [2] measure. This allow us to establish the relationship between two variables: X and Y.

By doing that, we must remember one simple rule: **Correlation does not imply causation.** When we are looking for a correlation, we are using a statistic tool that helps us to analyse and define a relation between two or more variables. This may lead us to a positive or negative correlation [3].

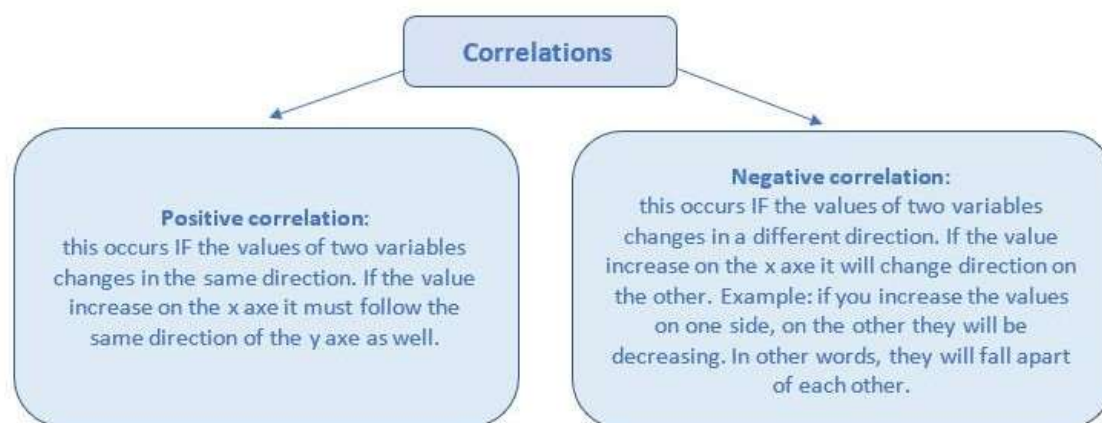


Figure n° 3: Difference between the two types of correlations. (Credits: create by the author).

We must follow some formulas to discover all the values we are going to need.

To do that I have to start recognizing some of its component in my assignment.

I'll start describing all the steps I've took since now, used into excel to develop all my assignment calculations.

Step-by-step development:

The process starts by dividing every column and assigning the proper name/letter to it.

The first column is the one that indicates the number of patients taken in consideration. I will name it as **N** during the calculation process. This will help me to identify how many patients have been taken onto consideration for the **population sample** [4] for each group, white and BAME.

Sampling methods:				
	Random Sampling:	Systematic Sampling:	Stratified Sampling:	Clustered sampling:
What is:	Each person that can be allowed to the sampling group is randomly chosen and everyone has the same chances to be selected.	The individuals selected for this sampling are chosen following regular intervals. Those ensure that an appropriate size of sampling subjects is taken under consideration.	The population is divided into groups which have related characteristics. It's commonly applied when is expected to have measures of interest which varies among the subgroups.	Instead of using individuals, it divides the population onto subgroups/clusters randomly selected and used for the study.
Advantage:	It's the best logical for probability sampling.	Is easy to manage.	It improves the accuracy and lower the probabilities of favouritisms which allows to represent the results in an easier way.	It may be more effective than random sampling.
Disadvantage:	It might happen that the subjects selected for this sampling may not have enough characteristics in common.	It may lead to favouritism	It needs knowledge about the right characteristics for the sample groups and is not easy to decide between the characteristics to be used.	High risk of prejudice.

Figure n° 4: sampling methods. (Credits: created by the author).

In the second column, proceeding to the right, I'll find the first group which consist of the White patients observed, I will name that as X.

In the third column, I'll find the BAME patients observed, this, will be renamed as Y.

I'm now starting to develop the process looking to find (if any) a relationship between my X and Y, following the Pearson Correlation Coefficient.

I will list the followed steps below:

1. I'm starting by establishing the Mean values of both, X and Y. Those will be calculated by doing the Summation of X and Y (Σ_X ; Σ_Y) by adding every single number below every corresponding column. Once obtained Σ_X and Σ_Y , to obtain the mean for each one, which will be written as: $\bar{x}$ and $\bar{y}$ from now, I will have to take the total of the summation and to divide this number by N ($\bar{x} = \frac{\Sigma_X}{N}$; $\bar{y} = \frac{\Sigma_Y}{N}$).
2. Once I've done this, I can start calculating the value of x and y, which comes out by doing the following: $X - \bar{x}$ and $Y - \bar{y}$, for each value faced in every column. I will start writing down the respective values onto a new column, named respectively, x and y.
3. After that, I must find out the xy value, which is equal to each value of x multiplied by each value of y [$(x) \cdot (y)$], following a line-by-line order.
4. I've then calculated the summation of xy (Σ_{xy}) and then used it to estimate the value of the covariance of xy (cov x y), by dividing the summation of xy by N-1 ($\text{cov } x y = \frac{\Sigma_{xy}}{N-1}$).
5. I've then calculated x and y at the power of two (x^2 ; y^2); still following line after line, value after value. And then, I've computed the respective summations of x and y at the power of two (Σ_{x^2} ; Σ_{y^2}).
6. Once I had obtained those values, I was able to estimate the values of the variance of x and y, by using this formula: $\text{var } x = \frac{1}{N-1} \cdot \Sigma_{x^2}$; $\text{var } y = \frac{1}{N-1} \cdot \Sigma_{y^2}$
7. Once having done so, we have just a few more steps to go. I must calculate the standard deviations of x and y by following these simple formulas: $SDx = \sqrt{\text{var } x}$; $SDy = \sqrt{\text{var } y}$. Which means that to obtain the value of the standard deviation of x, I will have to calculate the square root of the variance of x; and for the value of the standard deviation of y I will have to calculate the square root of the variance of y.
8. Once I have all those values, I will then be able to calculate the value of r_{xy} which determines the coefficient of the correlation between x and y.

To do so, I must follow this formula: $r_{xy} = \frac{\text{cov } xy}{SDx \cdot SDy}$

Anyway, this operation helped us just to identify the relation between the two variables but nothing more.

If we want to know if the y is increasing, and why and how much, we must proceed with Linear Regression [5].

This establishes the impact that a change on an independent variable has on a dependent variable.

Dependent variables, correlation, and linear regression:

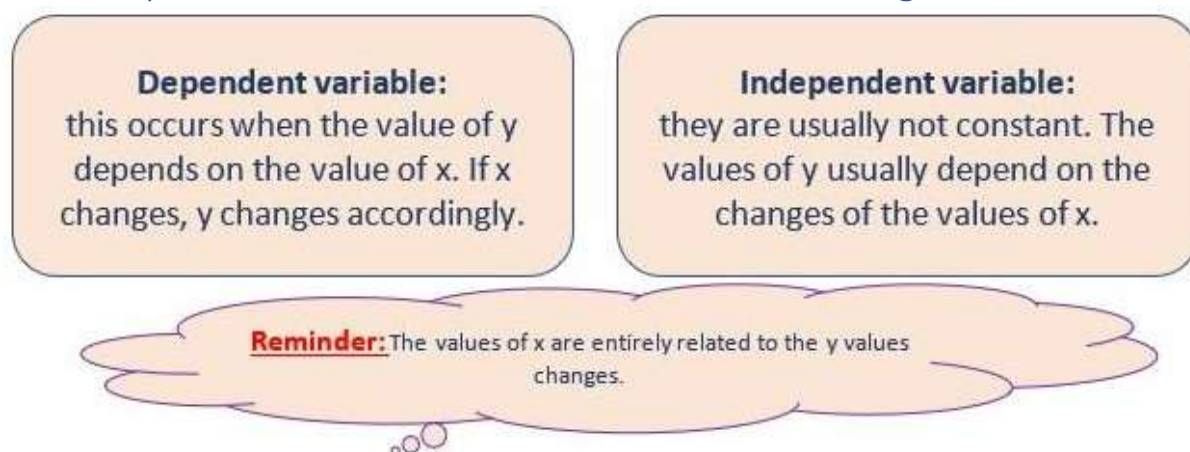


Figure n° 5: Difference between Dependent and Independent variables. (Credits: created by the author).

This, it's a statistical method that try the data mind, the strength and the character of relationship between Dependent variable (y) and Independent variable (x).

An easy and good reminder for correlation and linear regression is:

If x values changes, y values changes accordingly.

Linear Regression [6] helps us to predict the values of the changes.

However, that was just a small bracket related to what we could need to produce some other hypothesis that I will not apply in this case.

Standard error. What is and how to predict when it occurs:

Based on the knowledge of Linear regression and correlation, we now must consider the possibility to occur into errors during our predictions.

That brings us to the concept of **STANDARD ERROR** [7].



Figure n° 6: Standard error types. (Credits: created by the author).

Ways to understand which type of error is occurring:

Examiner's decision	H_0 is TRUE	H_0 is FALSE
Chooses to reject the H_0	<u>Type 1 error is occurring.</u>	It is the right decision to be taken.
Chooses to accept the H_0	It is the correct decision to be taken.	<u>Type 2 error is occurring.</u>

Figure n° 7: Type of standard error occurring. (Credits: created by the author; References: Week 7 power point slide, Standard Error explanation Dr. Anupam Mazumdar).

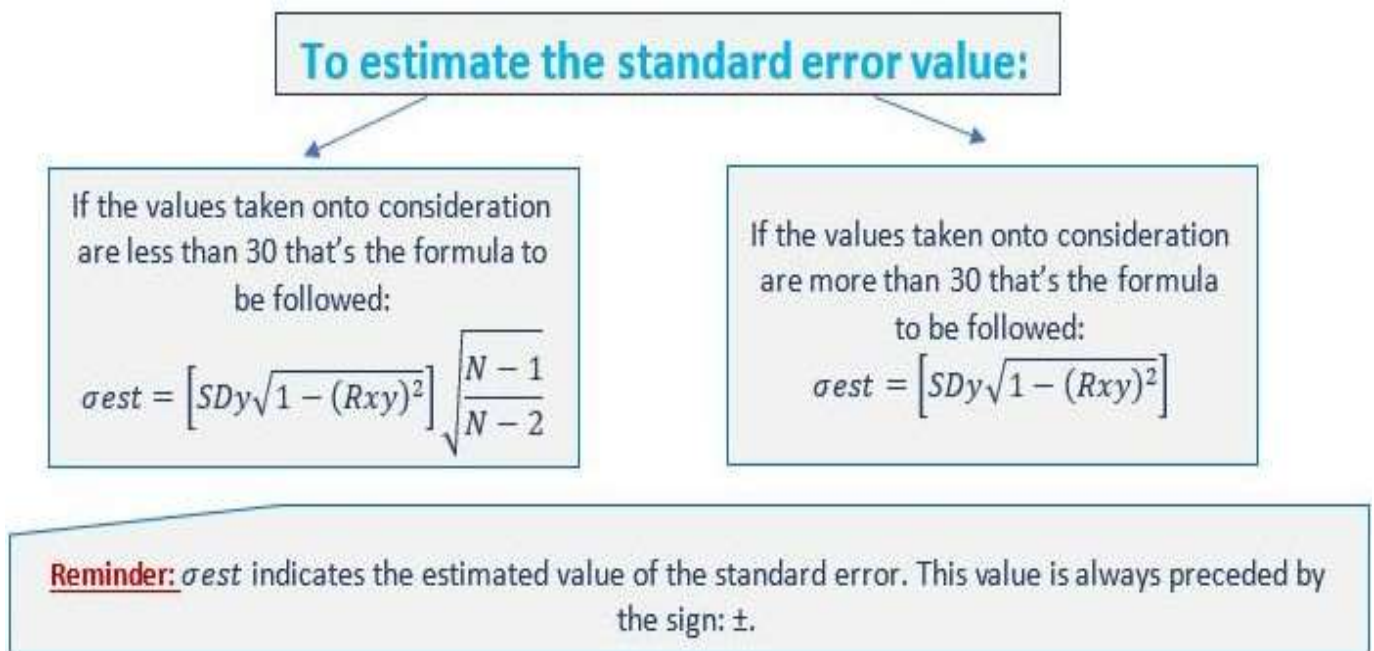


Figure n° 8: formula to estimate the value of the standard error. (Credits: created by the author).

Null Hypothesis and Alternative Hypothesis:

Thanks to the value obtained as a result onto the r_{xy} we can establish if the grade of the correlation is strong or weak, comparing this result to the **pvalue** (value of the probability) which usually is equal to 0.05.

This allows us to choose to accept or reject the null hypothesis [8] [9].

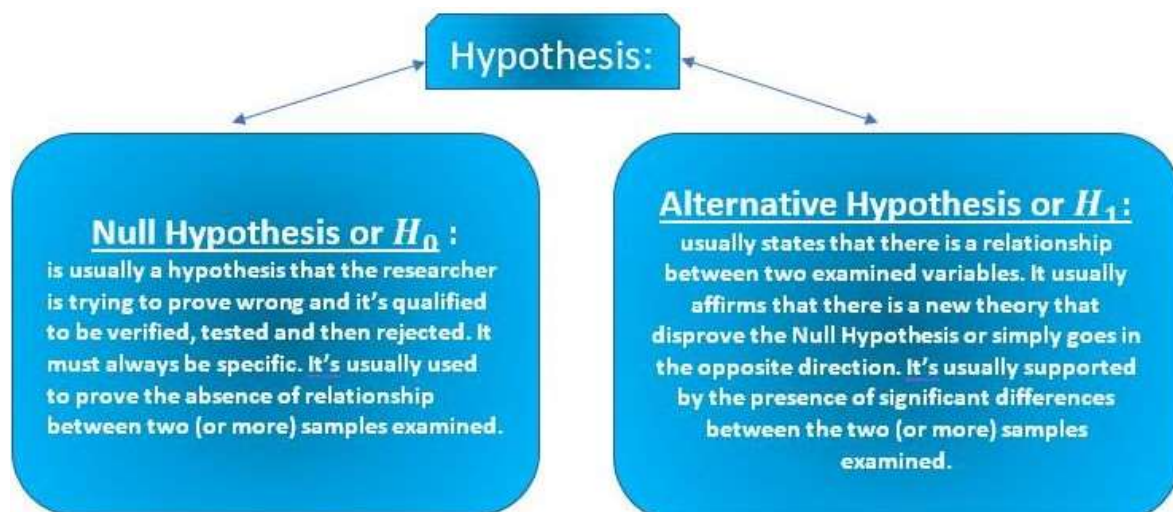


Figure n° 9: Null Hypothesis and Alternative Hypothesis, and how to determine which one to use. (Credits: created by the author).

T-test. What is and what is used for:

Thanks to this, and observing my assignment, I have now left to understand and then choose which t test I will have to apply to this data set, and I will then evaluate all the possibilities in here.

But before doing that, we should understand better what exactly a t-test [10] is and what do we need it for.

A **t-test** is commonly used to determine where and if there are significant differences between two (or more) different data sets considering two (or more) populations samples. Depending on the difficulties faced to identify the mean of the populations we have then to choose between the equal or unequal t-test.

If we must consider the mean of the population samples, we will then have to choose between the z-test or the t-test for two paired means sample.

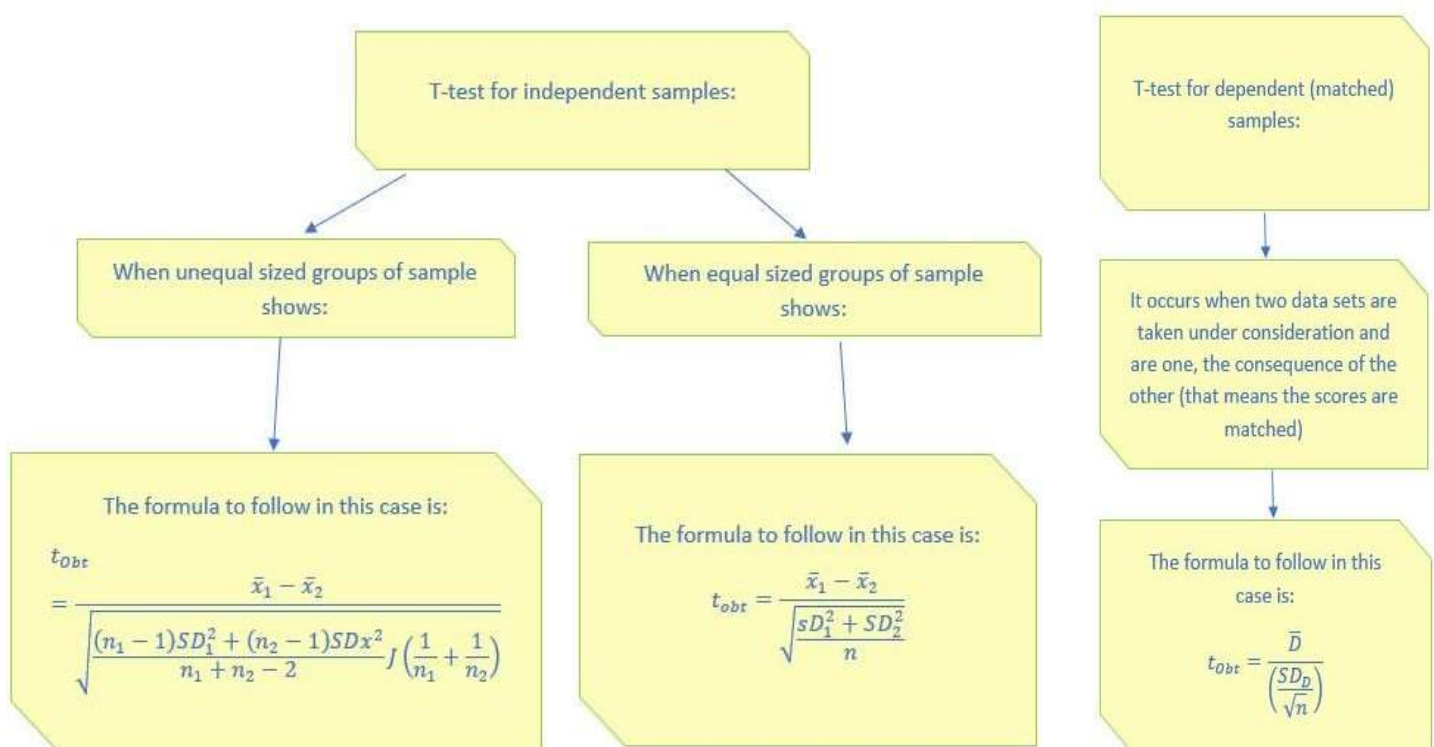


Figure n° 10: Types of t- test and its relative formulas to be followed. (Credits: created by the author).

The **t_{obt} value** is vital for the next steps into my assignment.

Based on its result, I can now check if the sample considered are statistically significant.

To do so I will need other three important things:

- First, I must find what's called the **t_{crit} value**. To do so, I must have easy access to the table that follows and to two additional information.

df	One-Tailed Tests			Two-Tailed Tests		
	.10	.05	.01	.10	.05	.01
1	3.078	6.314	31.821	6.314	12.708	63.857
2	1.886	2.920	6.965	2.920	4.303	9.925
3	1.638	2.353	4.541	2.353	3.182	5.841
4	1.533	2.132	3.747	2.132	2.776	4.604
5	1.476	2.015	3.365	2.015	2.571	4.032
6	1.440	1.943	3.143	1.943	2.447	3.708
7	1.415	1.895	2.998	1.895	2.365	3.500
8	1.397	1.860	2.897	1.860	2.306	3.356
9	1.383	1.833	2.822	1.833	2.262	3.250
10	1.372	1.813	2.764	1.813	2.228	3.170
11	1.364	1.796	2.718	1.796	2.201	3.106
12	1.356	1.783	2.681	1.783	2.179	3.055
13	1.350	1.771	2.651	1.771	2.161	3.013
14	1.345	1.762	2.625	1.762	2.145	2.977
15	1.341	1.753	2.603	1.753	2.132	2.947
16	1.337	1.746	2.584	1.746	2.120	2.921
17	1.334	1.740	2.567	1.740	2.110	2.898
18	1.331	1.734	2.553	1.734	2.101	2.879
19	1.328	1.729	2.540	1.729	2.093	2.861
20	1.326	1.725	2.528	1.725	2.086	2.846
21	1.323	1.721	2.518	1.721	2.080	2.832
22	1.321	1.717	2.509	1.717	2.074	2.819
23	1.320	1.714	2.500	1.714	2.069	2.808
24	1.318	1.711	2.492	1.711	2.064	2.797
25	1.317	1.708	2.485	1.708	2.060	2.788
26	1.315	1.706	2.479	1.706	2.056	2.779
27	1.314	1.704	2.473	1.704	2.052	2.771
28	1.313	1.701	2.467	1.701	2.049	2.764
29	1.312	1.699	2.462	1.699	2.045	2.757

(Continued)

190

TRWITYSTWDYDYMBAKHUQKUNZLZG2HWPIA3DQZMIRKALIMJENZ*Q5=

Appendix B - Critical Values of t 191

30	1.311	1.696	2.458	1.698	2.043	2.750
35	1.306	1.690	2.438	1.690	2.030	2.724
40	1.303	1.684	2.424	1.684	2.021	2.705
45	1.301	1.680	2.412	1.680	2.014	2.690
50	1.299	1.676	2.404	1.676	2.009	2.678
55	1.297	1.673	2.396	1.673	2.004	2.668
60	1.296	1.671	2.390	1.671	2.001	2.661
65	1.295	1.669	2.385	1.669	1.997	2.654
70	1.294	1.667	2.381	1.667	1.995	2.648
75	1.293	1.666	2.377	1.666	1.992	2.643
80	1.292	1.664	2.374	1.664	1.990	2.639
85	1.292	1.663	2.371	1.663	1.989	2.635
90	1.291	1.662	2.369	1.662	1.987	2.632
95	1.291	1.661	2.366	1.661	1.986	2.629
100	1.290	1.660	2.364	1.660	1.984	2.626
∞	1.282	1.645	2.327	1.645	1.960	2.576

Figure n° 11: Distribution table. (Credits: Statistics for the terrified, Kranzler, John H.).

- To be able to estimate the t_{crit} value, I must first find the df (Degree of Freedom), which is obtained by computing: $n_1 + n_2 - 2$ (in this case).
There are some exceptions. It could happen to have two different datasets, apparently not related; if examined/compared they come out to be one the result of the other. In this case, instead of computing two times the n values as once for each column, we just consider it as a total valid for both the groups, so the formula will be: $n - 1$
- I also need to know the α value or p value (which usually is specified into the exercises).

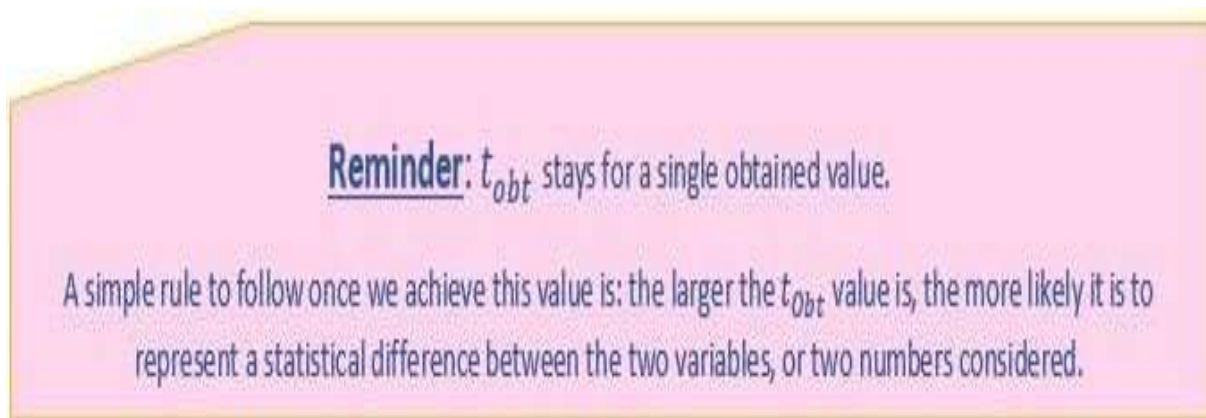


Figure n° 12: reminder for t_{obt} value. (Credits: created by the author).

Type of t-test run, motivations and analysis of the results outputted:

I've chosen to run the "t-Test: Two-Sample Assuming Unequal Variances" in Excel, and I've developed the step-by-step evaluation following the formula for the independent samples t-test considering equal sized samples (first figure in the assignment).

That was Excel's output:

t-Test: Two-Sample Assuming Unequal Variances		α value =0.10
	Variable 1	Variable 2
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	

Figure n° 13: t-test two sample, assuming Unequal variances with p value=0.10. (Credits: created by the author using Microsoft Excel).

Even though the N value observed is equal, the variables observed are **independent**.

The μ values for Variable1, and 2, doesn't seem to be different between each other, however we can identify a small difference, which can indicate a slightly higher probability of contagion rate onto the BAME group.

Following all the above stated rules, we can see that the observations taken under consideration are 50 on both sides: White (X) and BAME (Y). That leads us to apply the following formula to find the df :

$$n_1 + n_2 - 2$$

which gives us the result of 98, as we can see above. 100 words

We now must consider that our **significant value** is = 0.10,

and our **tStat value** is =1.66

and follow this rule:

IF	Significant level value (α value)	We can then:
Is higher than the p value	The t-test has a significant value/ The t-test is significant	Reject the Null Hypothesis or H_0
Is lower than the p value	The t-test has not significant value/ The t-test is not significant	Accept the Null Hypothesis or H_0

Figure n°14: Significant level value rules. When to accept or reject a H_0 . (Credits: created by the author).

Accepting or rejecting the Null Hypothesis:

Accordingly, to that rule above, observing the t_{crit} value and comparing it to the t_{stat} value, we can reject the H_0 .

Considering that:

$$t_{crit \text{ two tail}} > t_{stat}$$

we can say that the test is significant and when a significant difference occurs, we must then reject the H_0 .

We now must compare, following the same rule, the $p_{two - tail}$ value, with the $p_{one - tail}$ value. The result is the same:

$$p_{two - tail \text{ value}} > p_{one - tail \text{ value}}.$$

T-test run with significant value of 0.10:

Let's analyse the table above, with the $\alpha \text{ value} = 0.10$:

- $\alpha \text{ value} > t_{Stat \text{ value}} \longrightarrow 0.10 > -0.29$
that brings us to the conclusion that the test is significant
- $t_{crit \text{ two tail}} > t_{stat} \longrightarrow 1.66 > -0.29$
In this case the test is significant.
- $p_{two - tail \text{ value}} > p_{one - tail \text{ value}} \longrightarrow 0.76 > 0.38$

In relation to this, we can say once more, that the test has a significant value; so, we must reject the H_0 .

T-test run with significant value of 0.01:

t-Test: Two-Sample Assuming Unequal Variances		$\alpha \text{ value}=0.01$
	Variable 1	Variable 2
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	

Figure n° 15: T-test two sample, assuming Unequal variance with p value=0.01. (Credits: created by the author).

Considering what we stated before, we can now analyse and say that if the $\alpha \text{ value}=0.01$,

- $\alpha \text{ value} > Stat \text{ value} \longrightarrow 0.01 > -0.29$
Conclusion: the test is significant.
- $t_{crit \text{ two tail}} > t_{stat} \longrightarrow 2.62 > -2.36$
Conclusion: the test has significant value.

$$\text{➤ } p \text{ two-tail value} > p \text{ one-tail value} \longrightarrow 0.76 > 0.38$$

In relation to this, we can say once more, that the test has a significant value; so we must reject the H_0 .

T-test run with a significant value of 0.02

t-Test: Two-Sample Assuming Unequal Variances

α value=0.02

	Variable 1	Variable 2
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	

Figure n° 16: T-test two sample, assuming Unequal variance with p value=0.02. (Credits: created by the author).

Analysing the last hypothesis table with α value = 0.02:

- $\alpha \text{ value} > \text{Stat value} \longrightarrow 0.02 > -0.29$
Conclusion: the test is significant.
- $t \text{ crit two tail} > t_{\text{stat}} \longrightarrow 2.36 > -2.08$
The test has significant value.
- $p \text{ two-tail value} > p \text{ one-tail value} \longrightarrow 0.76 > 0.38$

In relation to this, we can say once more, that the test has a significant value; so, then we must reject the H_0 .

Null Hypothesis and Alternative Hypothesis regarding this test:

It's now time to build up our Null Hypothesis.

Knowing the previously cited rules to build a H_0 , my hypothesis will be the following:

- ✚ H_0 : The rate of infection between BAME and White people groups is varied
- ✚ H_1 : The rate of infection between BAME and White people groups is unvaried or equal

Considering that, and all the values studied before, if we want to avoid a type 2 error, we must reject the H_0 . However, we must consider that there is a **prejudice** into the white patients' sample caused by a higher age (55 years) than the range considered.

That force us to face a Type 2 error and to accept the H_0 .

To prove my choice, I'll leave as follow a graph which represent the tendency of the rate of infection between the two groups.

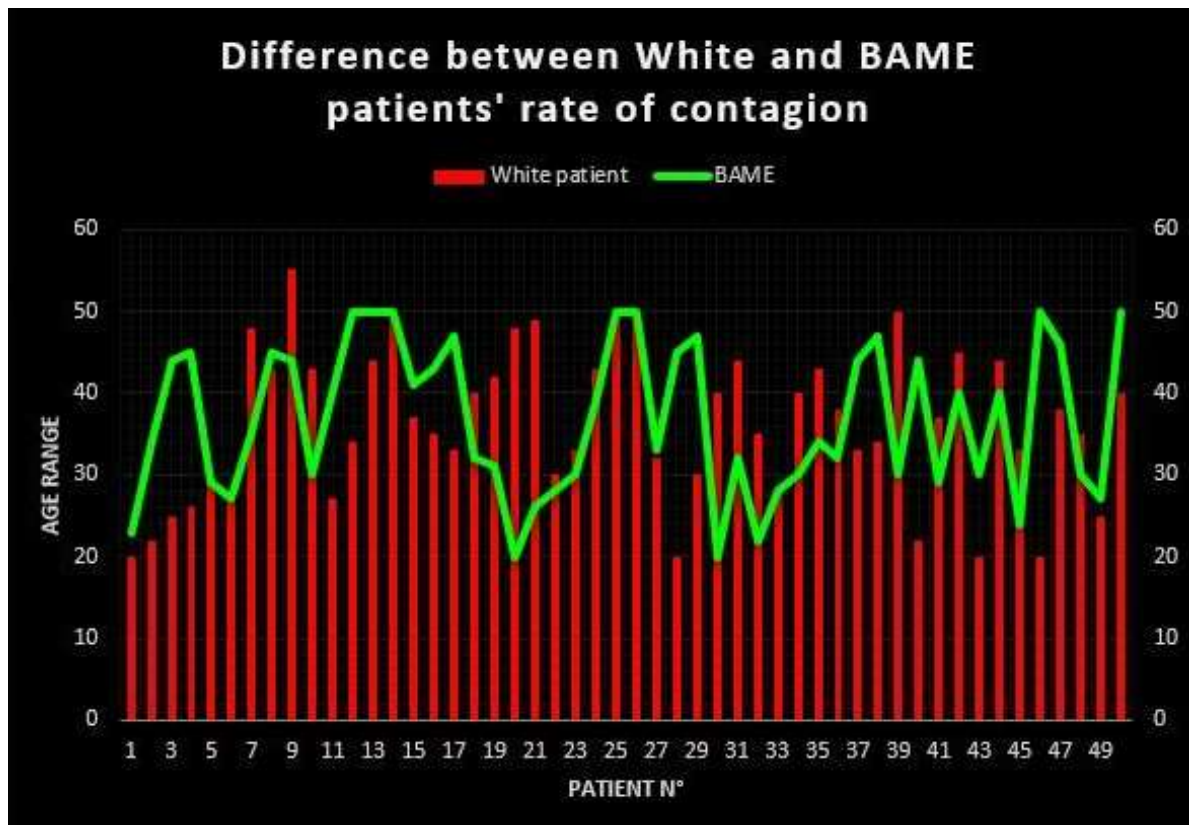


Figure n° 17: Infographic showing the differences between White and BAME patients' rate of contagion. (Credits: created by the author).

Visualisation of the age range and frequency of the infection between the age ranges

We will now visualise the data divided by age and frequency of the infection ratio in the graph:

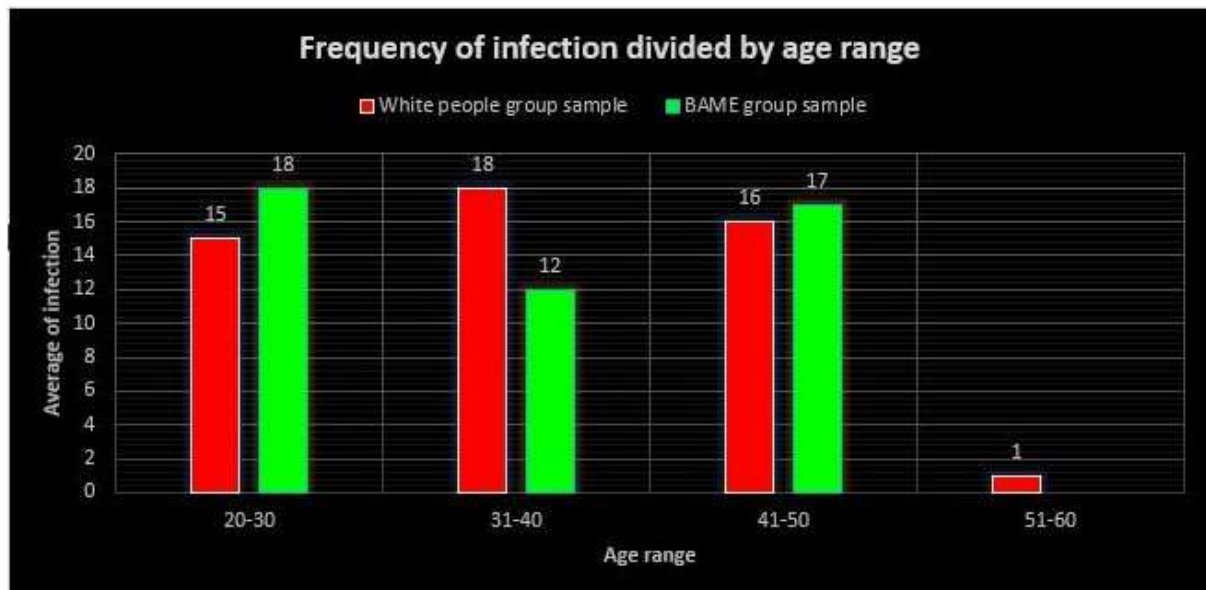


Figure n° 18: Infographic showing the frequency of infection rate divided by age range. (Credits: created by the author).

We can here see, that for the first age range 20-30 the infection ratio is higher into the BAME community. The situation is inverted if we take a closer look to the 31-40 age range. While onto the 41-50 range they are almost the same. The study could have had a deeper impact if who were collecting the date had taken under consideration the sex of the patients as well. In such a case, the observation could have been more specific.

If we want to make the observation further specific, we must consider, some other factors such as:

- ✚ Everyday life habits
- ✚ Precaution taken
- ✚ Physical pathologies
- ✚ Health problems
- ✚ Employment sector
- ✚ Range of contact with other people/family members outdoor
- ✚ Family composition and age range indoor
- ✚ Previous ongoing infections

Conclusion:

Considering all the observations, if we choose to follow what the rules say and the values obtained, we **must reject** H_0 saying that the rate of infection is the same and unvaried. However, observing the graphs and the marginally differences faced, I choose to face a type 2 error, affirming that the infection ratio is varied between the groups examined. However, we can't be 100% sure that our observation or the hypothesis are going to be right. To make the research more specific, we should stratify in a deeper way the observations considering even other features.

References

- [1] M. R. guy, "My market research methods," 05 10 2020. [Online]. Available: <https://www.mymarketresearchmethods.com/descriptive-inferential-statistics-difference/>. [Accessed 20 12 2020].
- [2] j. Magiya, "towards data science," 27 05 2019. [Online]. Available: <https://towardsdatascience.com/pearson-coefficient-of-correlation-explained-369991d93404>. [Accessed 15 12 2020].
- [3] L. Thomas., "Scribbr," 20 05 2020. [Online]. Available: <https://www.scribbr.com/methodology/independent-and-dependent-variables/>. [Accessed 07 12 2020].
- [4] H. Barratt, "Health knowledge," 2009. [Online]. Available: <https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/methods-of-sampling-population#:~:text=In%20this%20case%20each%20individual,decide%20which%20individuals%20to%20include..> [Accessed 05 11 2020].
- [5] S. Swaminathan, "towards data science," 26 02 2016. [Online]. Available: <https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86>. [Accessed 24 11 2020].
- [6] R. Bevans, "Scribbr," 28 01 2020. [Online]. Available: <https://www.scribbr.com/statistics/statistical-tests/>. [Accessed 20 12 2020].
- [7] I. P. J., "Hypothesis testing, type I and type II errors," *Hypothesis testing, type I and type II errors*, no. 2009, p. 131, 2009.
- [8] S. McCombes, "Scribbr," 23 04 2019. [Online]. Available: <https://www.scribbr.com/research-process/hypotheses/>. [Accessed 29 11 2020].
- [9] A. Sapkota, "Microbe Notes," 23 05 2020. [Online]. Available: <https://microbenotes.com/null-and-alternative-hypothesis/>. [Accessed 02 12 2020].
- [10] W. Kenton, "Investopedia," 22 03 2020. [Online]. Available: <https://www.investopedia.com/terms/t/t-test.asp>. [Accessed 26 12 2020].