



Data Analysis and Visualization

DAV (ACCB4002)



DAV, Assessment 2

Aleksandra Markova 2011971

Lecturer: Anupam Mazumdar

What is this T -test?

A T-test is used to find out if two datasets are significantly different from each other. It compares the two sets of data and their means. The test assumes the data follows a normal distribution and two groups of data to be equal. The test uses Null Hypothesis and Alternative Hypothesis. The Null Hypothesis states there is no difference between the groups. On another side, the Alternative Hypothesis states difference exists between the groups. The value resulting from the T-test is called a p-value. It tells us how likely the difference between the two groups could have happened. If the p-value is less than 0,05-called critical value, then there is a statistically significant difference. But before I get to compare p-value with the critical value there are some calculations that I will be needed. Three main values are needed for this purpose: mean difference, standard deviation, and amount of data of each group. I am going to explain every single step how to get to the P-value in my task.

The development of an appropriate set of testable hypotheses for the experiment:

In my task in which is given must identify, that the infection rate of Covid-19 is varied between white people and BAME in the United Kingdom and need to identify the ratio and the truthfulness. For this purpose, I was given 50 different white and BAME patients. All patients are between the ages of 20 and 50. In first place a need to find out the mean value ($\bar{x}$). For this purpose, I need the summation of the age of all the patients, as follows: white people- 1809 ($\bar{X}$) and BAME- 1837 ($\bar{Y}$).

$\bar{X} = \sum X/n$ (where n is a number of patients). $\bar{X} = 1809/50 = 36.18$
 $\bar{Y} = \sum Y/n$; $\bar{Y} = 1837/50 = 36.74$. The results: **36.18** and **36.74** are called 'mean'. I can see that the two means are almost the same but still cannot tell do I reject or cannot reject the Null Hypothesis. To find out standard deviation first I need a variance, which I can use with this formula: $\sigma^2 = \sum (x - \bar{x})^2 / n$; and for the second group $\sigma^2 = \sum (y - \bar{y})^2 / n$. With those formula I found variances for white people 88.72204082 and BAME 86.80857143. Now to find out the standard deviation must square root those values. For the first one: $\sqrt{88.72204082} = 9.418284$, and for the second one: $\sqrt{86.80857143} = 9.317111$

9.418284 $\searrow$ standard

9.317111 $\nearrow$ deviation

The t-test delivers two values at the end: Degree of freedom, which values it has the freedom to vary and are important for measuring the importance and the strength of the null hypothesis. The degree of freedom formula is the number of observations minus one but because in my task there are two different observation is minus 2 or: $n - 2$, or in the given task two different groups of 50 white people and 50 BAME, which mean $50 + 50 = 100 - 2 = 98$. The degree of freedom in my task is **98**. The second value is t-value. Bigger t- value shows there is a big difference exists between the two groups. Smaller, they are almost the same. I can find it out when find difference between group

means or $(\bar{x} - \bar{y})$ divided by variability of the groups I will use the factor $\sqrt{S_x^2/n_x + S_y^2/n_y}$. 'S' is a standard deviation but actually is squaring that that gives me a variance. I am going to apply the formula with the values which I reached so far. $(36.18 - 36.74) / \sqrt{88.72204082/50 + 86.80857143/50} = -0.56 / \sqrt{1.77444082 + 1.73617143} = -0.56 / \sqrt{3.51061225} = -0.56 / 1.87366279 = -0.29887982$. Now once I know the t- value I need to compare with a value obtained from a critical value table (T- Distribution table). This is going to help to clarify the cause of possibilities on the difference does exist or the difference is outside that chance to vary. The T- Distribution Table is presented in one tail and two tails test. The first one is statistical test, when the critical value of distribution is bigger or smaller than a specific value but not all together. Two tails test when the critical value of distribution is bigger or smaller than a specific variety of values. It is used in Null Hypothesis and for statistical meaning testing.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.90}$	$t_{.95}$	$t_{.98}$	$t_{.99}$	$t_{.995}$	$t_{.998}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002
df										
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.960	2.330	2.581	3.098
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%
										99.9%

Confidence Level

Image by the author

This is how looks like the t- table. Since the value of degree of freedom is not in the table (98) I use the closest which in this case is 100. Comparing with one tail value is 1.660 and with two tail value is 1.984. We observed the value 0.05 also called alpha value or significance level. In most cases I use 0.05 or 5% as significant level but also can use other alphas (I will show how later). This alpha value exactly means that if I do this comparing between these two groups a hundred times 95 of the times, I will reject the Null Hypothesis and only 5% I would not. To compare already obtained t- value I will use the two-tail column: 1.984. The rule is: **if t- value is smaller than t- critical two tail value: cannot reject the Null Hypothesis. And if t- value is bigger than t- critical two tail value: can reject the Null Hypothesis.** The t- value is -

0.29887982 and is smaller than t- critical two tail value which is **1.984**, therefore I cannot **reject the Null Hypothesis**. Now is the moment to do the T- test and to find the p- value which also must compare with the t- critical two tail value.

Now using this t- table I can find out the approximately where is the p- value. The t critical value is - **0.29887982**. I am going to ignore the sign 'minus' and now I am looking for number in the **0.29887982** on the table row 100 because the degree of freedom is 98.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%

Confidence Level

Image by the author

That means the p- value is in one tail alpha bigger 0.25 or smaller 0.50 (0.25 < p value < 0.50), or in two tail alpha: bigger then 0.50 and smaller 1 (0.50 < p value < 1). The exact P value I can calculate with Microsoft Excel.

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	36.18	36.74
Variance	88.72204082	86.80857
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	

Graphic 1

In the Excel table I can see that the p- value one tail is 0. 382832305, and two tail p- value is 0. 76566461 (which is exactly one tail multiply by 2). And both values agree to the statements from above.

one tail: $0.25 < 0.382832305 < 0.50$

two tail: $0.50 < 0.76566461 < 1$

For my Hypothesis I am going to use two tail p- value for comparison, which means the p- value is bigger or smaller than 0,05. **Therefore 0.76566461 > 0.05. In this case I cannot reject the Null Hypothesis.**

There are two types of T- tests. Equal and Unequal T-test. Equal is used when the dataset is small and can easily identify their mean value. Unequal is used when the dataset is not relatively small and difficult to identify their means. In my task I must use Unequal T- Test because in the given dataset is not small enough to find out the mean value. I am going to show example for Equal t-test:

Student	Class A	Class B
1	147	700
2	547	256
3	465	512
4	863	150
5	356	107
6		546
7		675

Graph 2

In this graph 2 it is easy to see when I must use Unequal t- test. The number of the observations are with unequal number observation.

Testing of my hypothesis to significant level of $p=0.10$, $p=0.01$, $p=0.02$

What is Statistically Significant?

'Statistical significance refers to the claim that a result from data generated by testing or experimentation is not likely to occur randomly or by chance but is instead likely to be attributable to a specific cause. Statistical significance can be considered strong or weak. When analysing a data set and doing the necessary tests to discern whether one or more variables have an effect on an outcome, strong statistical significance helps support the fact that the results are real and not caused by luck or chance. Simply stated, if a p-value is small then the result is considered more reliable.' [1]

Now after I used the significant level 0.05 in the t -test, I am going to test in the hypothesis level $p=0.10$. For that purpose, I am using the same data, same difference between the mean. The only value which I have changed is the alpha level from 0.05 with 0.10.

3		Variable 1	Variable 2
4	Mean	36.18	36.74
5	Variance	88.72204082	86.80857
6	Observations	50	50
7	Pooled Variance	87.76530612	
8	Hypothesized Mean	0	
9	df	98	
10	t Stat	-0.298879822	
11	P(T<=t) one-tail	0.382832305	
12	t Critical one-tail	1.290249904	
13	P(T<=t) two-tail	0.76566461	
14	t Critical two-tail	1.660551217	

Graphic 3

As I can see the different values are the t critical one and two tail. There is not a significant difference between them but still there is. Or $1.6605 > 0.29887$ as well as $0.76566 > 0.10$. In this case I still can claim that **I cannot reject the Null Hypothesis**.

Next, I will test significant level $p=0.01$.

3		Variable 1	Variable 2
4	Mean	36.18	36.74
5	Variance	88.72204082	86.80857
6	Observations	50	50
7	Pooled Variance	87.76530612	
8	Hypothesized Mean	0	
9	df	98	
10	t Stat	-0.298879822	
11	P(T<=t) one-tail	0.382832305	
12	t Critical one-tail	2.36500241	
13	P(T<=t) two-tail	0.76566461	
14	t Critical two-tail	2.626931096	
15			

Graphic 4

Here T critical two tail is bigger than t value: $2.626931 > 0.298879$, also p value is bigger than alpha: $0.76566 > 0.01$, which also prove that **I cannot reset the Null Hypothesis**.

My last testing will be with significant level $p=0.02$.

3		<i>Variable 1</i>	<i>Variable 2</i>
4	Mean	36.18	36.74
5	Variance	88.72204082	86.80857
6	Observations	50	50
7	Pooled Variance	87.76530612	
8	Hypothesized Mean Dif	0	
9	df	98	
10	t Stat	-0.298879822	
11	P(T<=t) one-tail	0.382832305	
12	t Critical one-tail	2.081444921	
13	P(T<=t) two-tail	0.76566461	
14	t Critical two-tail	2.36500241	

Graphic 5

Like others graphics the t critical two tail value is bigger than t value ($2.365 > 0.2988$), moreover p value is bigger than alpha ($0.7656 > 0.02$), and in this testing once **again I cannot reset the Null Hypothesis.**

The presence of any outliers in the data and the effect that they have on the outcome:

3		<i>Variable 1</i>	<i>Variable 2</i>
4	Mean	36.18	36.74
5	Variance	88.72204082	86.80857
6	Observations	50	50
7	Pooled Variance	87.76530612	
8	Hypothesized Mean Difference	1	
9	df	98	
10	t Stat	-0.832593789	
11	P(T<=t) one-tail	0.203549145	
12	t Critical one-tail	1.660551217	
13	P(T<=t) two-tail	0.40709829	
14	t Critical two-tail	1.984467455	

Graphic 6

In my hypothesizing that there is not a difference between the means, so I put 'zero' in the box Hypothesizing mean difference, but if I do think there is going to be a difference in the means, then I need to put the difference between the means. In graphic 6 I am showing different option. There I put in the box the number 1 and the t value now are completely different. Therefor the p value is different as well. All others value remaining the same.

In my task the last subtask is:

Any other factors that might affect the results but have not been considered in the stated experimental set-up, together with ideas as to how you might control these factors

There are a few factors:

- Depending how big is the difference between two groups. Even all other factors are same, bigger difference between the means I can be sure that there is statistically significant between the groups.
- Which alpha level is using in the test? Bigger alpha value means smaller difference between the means.
- The number of the objects in the two samples. The amount of the sample is very important in defining the significance of the difference between means.
- Similarity between the groups. When other factors are equal, smaller variances of the two groups under factor are the greater possibility that a statistically significant mean difference exists. I can be more convinced that two groups vary when the results within each group are close together.
- Am I testing one or two tail hypothesis?

Like a conclusion I will give this citation: *“Why we use t-test specifically and why not other tests?” Because in most of the process scenarios, we do not know the population parameters like mean or standard deviation and hence the only option is t-test [2]*

References

[1] S. Author, “Investopedia,” 2020.

[2] Qsutra, 2020.

Table of Contents

What is a T- Test	1
The development of an appropriate set of testable hypotheses for the experiment.....	1/3
Testing of my hypothesis to significant level of $p=0.10$, $p=0.01$, $p=0.02$	4/5
The presence of any outliers in the data and the effect that they have on the outcome: _	6
Any other factors that might affect the results but have not been considered in the stated experimental set-up, together with ideas as to how you might control these factors_	7
Reference.....	7