

INTRODUCTION TO WEB AND DATABASE CONCEPTS

ACCB4002 | CLOUD COMPUTING | SUNETH FERNANDO

ASSESSMENT 2 DATA ANALYSIS AND VISUALISATION

REPORT

SEPTEMBER 2021



PRESENTED BY

Plamena Gesheva

Student ID : 2019587



TABLE OF CONTENT

1.INTRODUCTION

2.T-TEST

- DEFINITION & UNDERSTANDING

3.HIPOTHESIS

- UNDERSTANDING & DEVELOPING

4. ANALYSIS- CONCLUSION

5. APPENDIX

6.REFERENCES



1. INTRODUCTION

Traditional methods for statistical analysis – from sampling data to interpreting results – have been used by scientists for thousands of years. But today's data volumes make statistics ever more valuable and powerful. Affordable storage, powerful computers and advanced algorithms have all led to an increased use of computational statistics.

Whether you are working with large data volumes or running multiple permutations of your calculations, statistical computing has become essential for today's statistician.

2.T-TEST

DEFINITION & UNDERSTANDING



A t-test is a statistic used to compare the averages of two groups (or challenge a hypothesis) to see if differences or similarities are real or just random chance. A t-test really is a way to use math to test relationships between groups — or to see if observed results match a hypothesis. The t-test looks at a set of data called the t-statistic, a measure of how well results fit a theory or expected result. Then it tries to figure out if differences — or similarities — between two sets of data are likely to happen all the time or if they were just random chance.[1]

EXAMPLE:

We want to know whether the mean petal length of iris flowers differs according to their species. You find two different species of irises growing in a garden and measure 25 petals of each species. You can test the difference between these two groups using a t-test.

2.T-TEST

DEFINITION & UNDERSTANDING



- The null hypothesis (H_0) is that the true difference between these group means is zero.
- The alternate hypothesis (H_a) is that the true difference is different from zero.

What type of t-test should I use?

When choosing a t-test, you will need to consider two things: whether the groups being compared come from a single population or two different populations, and whether you want to test the difference in a specific direction.

One-sample, two-sample, or paired t-test?

- If the groups come from a single population (e.g. measuring before and after an experimental treatment), perform a paired t-test.
- If the groups come from two different populations (e.g. two different species, or people from two separate cities), perform a two-sample t-test (a.k.a. independent t-test).

2.T-TEST

DEFINITION & UNDERSTANDING



- If there is one group being compared against a standard value (e.g. comparing the acidity of a liquid to a neutral pH of 7), perform a one-sample t-test.

Performing a t-test

The t-test estimates the true difference between two group means using the ratio of the difference in group means over the pooled standard error of both groups. We can calculate it manually using a formula (fig.1), or use statistical analysis software.

A larger t-value shows that the difference between group means is greater than the pooled standard error, indicating a more significant difference between the groups.

We can compare your calculated t-value against the values in a critical value chart to determine whether your t-value is greater than what would be expected by chance. If so, you can reject the null hypothesis and conclude that the two groups are in fact different.[2]

3.HIPOTHESIS UNDERSTANDING & DEVELOPING



What is hypothesis testing?

In statistics, we may divide statistical inference into two major parts. One is estimation and another is hypothesis testing. Before hypothesis testing we must know about hypothesis, so we can define hypothesis. A statistical hypothesis is a statement about a population which we want to verify on the basis of information which contained in a sample(fig2).

A statistical test mainly involves four steps:

- Evolving a test statistic
- To know the sampling distribution of the test statistic
- Selling of hypotheses testing conventions
- Establishing a decision rule that leads to an inductive inference about the probable truth.

Types of statistical hypothesis

- Null hypothesis
- Alternative hypothesis[3]

3. HIPOTHESIS UNDERSTANDING & DEVELOPING



Null hypothesis

A null hypothesis is a statement, which tells us that no difference exists between the parameter and the statistic being compared to it. According to Fisher, any hypothesis tested for its possible rejection is called a null hypothesis and is denoted by H_0 .

Alternative hypothesis

The alternative hypothesis is the logical opposite of the null hypothesis. The rejection of the null hypothesis leads to the acceptance of the alternative hypothesis. It is denoted by H_1 .

For example, with a coin-tossing experiment, the null and alternative hypothesis may be formed as,

- **H_0** : the coin is unbiased.
- **H_1** : the coin is biased.

3.HIPOTHESIS UNDERSTANDING & DEVELOPING



Depending on the population distribution, the statistical hypothesis are two types,

- **Simple hypothesis:** when a hypothesis completely specifies the distribution of the population, then the hypothesis is called a simple hypothesis.
- **Composite hypothesis:** when a hypothesis does not completely specify the distribution of the population, then the hypothesis is called a composite hypothesis.[3]

3.HIPOTHESIS UNDERSTANDING & DEVELOPING



Depending on the population distribution, the statistical hypothesis are two types,

- **Simple hypothesis:** when a hypothesis completely specifies the distribution of the population, then the hypothesis is called a simple hypothesis.
- **Composite hypothesis:** when a hypothesis does not completely specify the distribution of the population, then the hypothesis is called a composite hypothesis.[3]

4. ANALYSIS CONCLUSION



Ratio and tested the hypotheses:

The thing is that it turns out that the null hypothesis is rough, i.e. there is significance in a person's skin tone. The level of significance it is higher than 0.05 i.e. there is another factor that affects the results(fig.3)(fig.4)[fig.5).

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

There is no regularity in the data, which confuses the testing of hypotheses, for example the lack of other indicators such as gender and accompanying diseases.

I take under consideration the total number of patients and the results show that the BAME are more infected. Factors that would affect the outcome but are not taken into account, most importantly age, maybe gender and of course the specifics of patients, whether they are smokers and whether they have concomitant diseases.

5. APPENDIX



$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

fig.1 T-TEST FORMULA

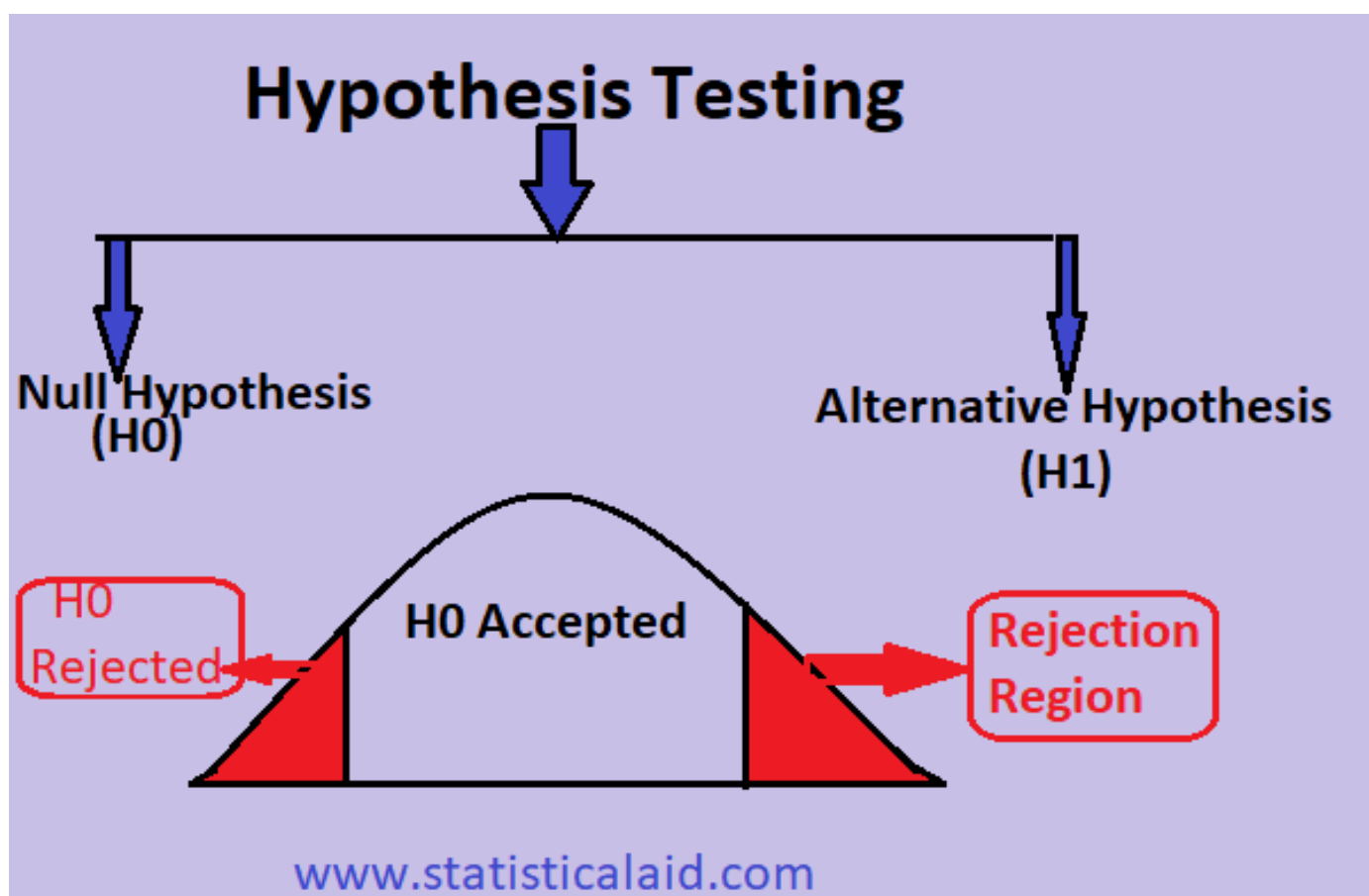


fig.2 HYPOTESIS TESTING

5. APPENDIX



H0			
H1			
F-Test Two-Sample for Variances			
	White Patient	BAME	
Mean	36.58	60.04	
Variance	143.5138776	547.182	
Observations	50	50	
df	49	49	
F	0.262278121		
P(F<=f) one-tail	3.35166E-06	hypothesis rejected	
F Critical one-tail	0.622165468		
Результат :			
F critical <= F < 1	H0 is accepted (equal dispersions)		
t-Test: Two-Sample Assuming Unequal Variances			
	White Patient	BAME	
Mean	36.58	60.04	
Variance	143.5138776	547.182	
Observations	50	50	
Hypothesized Mean Difference	0		
df	73		
t Stat	-6.312037424		
P(T<=t) one-tail	9.54795E-09		
t Critical one-tail	1.665996224		
P(T<=t) two-tail	1.90959E-08		
t Critical two-tail	1.992997126		

fig.3 Hypothesis Dataset given

5. APPENDIX



$F_{critical} \leq F < 1$	H_0 is accepted (equal dispersions)	H_1 rejected	$t_{stat} > t_{critical one-tail}$	H_0 is rejected in favor of the unilateral alternative H_1
$F \leq F_{Critical one-tail} < 1$	H_0 rejected (equal dispersions)		$t_{stat} \leq t_{Critical one-tail}$	H_0 се приема
$1 < F \leq F_{Critical one-tail}$	H_0 is accepted (equal dispersions)		$t_{stat} > t_{critical two tail}$	H_0 is rejected in favor of the unilateral alternative H_1
$1 < F_{Critical one-tail} \leq F$	H_0 rejected (equal dispersions)		$t_{stat} \leq t_{critical two tail}$	H_0 is accepted

fig.4 dataset given T-TEST

Layout	References	Mailings	Review	View	Help	RCM			
Patient No	White Patient	BAME	GDC	Num1	Num2	Ratio	Total	Ratio White patient	Ratio BAME
1	20	64		4	5	16 5:16	84	24%	76%
2	22	34		2	6	9 5.5:8.5	56	39%	61%
3	25	64		1	6	16 6.25:16	89	28%	72%
4	26	45		1	7	11 6.5:11.25	71	37%	63%
5	29	89		1	7	22 7.25:22.25	118	25%	75%
6	77	76		1	19	19 19.25:19	153	50%	50%
7	48	85		1	12	21 12:21.25	133	36%	64%
8	43	45		1	11	11 0.75:11.25	88	49%	51%
9	55	44		11	14	11 13.75:11	99	56%	44%
10	43	98		1	11	25 10.75:24.5	141	30%	70%
11	27	40		1	7	10 6.75:10	67	40%	60%
12	34	50		2	9	13 8.5:12.5	84	40%	60%
13	14	100		2	4	25 3.5:25	114	12%	88%
14	50	50		50	13	13 12.5:12.5	100	50%	50%
15	37	76		1	9	19 9.25:19	113	33%	67%
16	35	43		1	9	11 8.75:10.75	78	45%	55%
17	33	47		1	8	12 8.25:11.75	80	41%	59%
18	40	77		1	10	19 10:19.25	117	34%	66%
19	42	31		1	11	8 10.5:7.75	73	58%	42%
20	48	99		3	12	25 12:24.75	147	33%	67%
21	49	26		1	12	7 12.25:6.5	75	65%	35%
22	50	100		50	13	25 12.5:25	150	33%	67%
23	33	90		3	8	23 8.25:22.5	123	27%	73%
24	43	39		1	11	10 10.75:9.75	82	52%	48%
25	50	80		10	13	20 12.5:20	130	38%	62%
26	50	50		50	13	13 12.5:12.5	100	50%	50%
27	32	69		1	8	17 8:17.25	101	32%	68%
28	20	45		5	5	11 5:11.25	65	31%	69%
29	30	47		1	8	12 7.5:11.75	77	39%	61%
30	40	80		40	10	20 10:20	120	33%	67%
31	44	32		4	11	8 11:8	76	58%	42%
32	35	82		1	9	21 8.75:20.5	117	30%	70%
33	28	88		4	7	22 7:22	116	24%	76%
34	40	30		10	10	8 10:7.5	70	57%	43%
35	43	64		1	11	16 10.75:16	107	40%	60%
36	38	32		2	10	8 9.5:8	70	54%	46%
37	33	44		11	8	11 8.25:11	77	43%	57%
38	34	47		1	9	12 8.5:11.75	81	42%	58%
39	50	70		10	13	18 12.5:17.5	120	42%	58%
40	22	74		2	6	19 5.5:18.5	96	23%	77%
41	37	19		1	9	5 9.25:4.75	56	66%	34%
42	45	40		5	11	10 11.25:10	85	53%	47%
43	20	60		20	5	15 5:15	80	25%	75%
44	44	40		4	11	10 11:10	84	52%	48%
45	13	94		1	3	24 3.25:23.5	107	12%	88%
46	20	50		10	5	13 5:12.5	70	29%	71%
47	38	46		2	10	12 9.5:11.5	84	45%	55%
48	35	30		5	9	8 8.75:7.5	65	54%	46%
49	25	77		1	6	19 6.25:19.25	102	25%	75%
50	40	100		20	10	25 10:25	140	29%	71%
Total	1829	3002					4831	38%	62%

fig.5 Ratio Dataset given



6. REFERENCES

[1]

Robinhood. Ltd., “What Is a T-test?,”

<https://learn.robinhood.com/articles/>, Aug. 08, 2021.

<https://learn.robinhood.com/articles/f6z8jpZx64uk8xFLybwSs/what-is-a-t-test/> (accessed Aug. 16, 2021).

[2]

R. Bevans, “An Introduction to T-Tests | Definitions, Formula and Examples,” Scribbr, Jan. 31, 2020.

<https://www.scribbr.com/statistics/t-test/> (accessed Aug. 16, 2021).

[3]

S. Aid, “Statistical Hypothesis Testing step by step procedure - Statistical Aid,” Statistical Aid: A School of Statistics, Dec. 12, 2020. <https://www.statisticalaid.com/statistical-hypothesis-testing-step-by-step-procedure/> (accessed Aug. 31, 2021).