

DATA ANALYSIS AND VISUALISATION
ASSESSMENT 2

Lecturer:Anupam Mazumdar

Student:Pop Daciana Maria

Student ID:1919877

Executive summary

Data analysis and visualization is the process of providing a structure and order to the gathered data for representing a logical visualization of the collected dataset. This analysis of statistical data uses the BBC UK news data for the Covid-19 affected people infection rate analysis for white and BAME people. Analysis of the hypothesis test uses the statistical method t-test and correlation analysis process uses chi-square test method. SPSS software is used for developing both the statistical tests.

Table of Contents

Introduction.....	3
Analysis.....	3
Dataset description.....	3
Hypothesis development and analysis.....	4
Selection and justification of analyzing techniques.....	6
Selection and justification of techniques to present results.....	7
Application.....	7
Application of the techniques analysis.....	7
Interpretation of the result analysis.....	9
Presentation of results.....	10
Conclusion.....	10
Reference list.....	12

Introduction

This part of the data analysis and visualization focuses on the statistical analysis of the collected dataset of BBC news for the infection rate identification of Covid-19 in the country UK. To analyze and develop the selected dataset hypothesis and the ratio of infection rate this section uses t-test and chi-square test of statistical analysis by using SPSS software.

Analysis

Dataset description

Dataset is a collection of data that consists of one or more than one table and rows. All of the columns of the table represent a particular variable (Wendler, and Gröttrup, 2016, p. 33). To support the data analysis and data visualization concept of statistical data the BBC UK news dataset is used here. To analyze the ratio of infection rate and findings truthfulness of the Covid-19 Infection trends of the country UK, Covid-19 infection trend of white people and BAME patients of the UK is selected here. The selected dataset helps to identify 50 different patients and obtain their age of having the infection of the Covid-19 virus. The age of all the patients is between 20 years to 50 years. To analyze the characteristics of the selected Covid-19 infection age dataset of the BBC UK news one particular data table is selected here.

Patient No	White Patient	BAME	Patient No	White Patient	BAME
1	20	23	26	50	50
2	22	34	27	32	33
3	25	44	28	20	45
4	26	45	29	30	47
5	29	29	30	40	20
6	27	27	31	44	32
7	48	35	32	35	22
8	43	45	33	28	28
9	55	44	34	40	30
10	43	30	35	43	34
11	27	40	36	38	32
12	34	50	37	33	44
13	44	50	38	34	47
14	50	50	39	50	30
15	37	41	40	22	44
16	35	43	41	37	29
17	33	47	42	45	40
18	40	32	43	20	30
19	42	31	44	44	40
20	48	20	45	33	24
21	49	26	46	20	50
22	30	28	47	38	46
23	33	30	48	35	30
24	43	39	49	25	27
25	50	50	50	40	50

Figure1. Dataset

(Source: Provided)

The above data table is used for the data analysis and visualization of the UK Covid-19 infection rate data between white people and BAME people. This data table has five different columns where each of the columns has different variables. This dataset consists of 25 rows and each of the rows having five different values for five different attributes of the columns. All of the variables under each column consist of integer values. Those integer values are used for the statistical analysis purpose of identifying the ratio of infection rate and truthfulness of the findings.

Hypothesis development and analysis

Hypothesis development is a process of predicting some concept of value from the available evidence. To select the hypothesis application of statistical reasoning and output interest of the

provided dataset have to be considered. Hypothesis development helps to perform the t-test and chi-square test to analyze the statistical data and represent the visualization of the identified results (Yusup *et al.* 2018, p. 163). The confirmatory data analysis or the statistical hypothesis development approach helps to draw the interface from data. Regarding the selected dataset of the BBC UK news of Covid-19 infection rate in white people and BAME, the statistical hypothesis is going to be developed. A selective data portion is used for the t-test and chi-square test purpose by using the hypothesis development part. Different steps are required to develop the statistical hypothesis. According to this observation Null hypothesis develops when there is no difference between the infection rate of the white people of the UK and BAME of the UK. The Null hypothesis provided the idea where there are no variables of the dataset is related to each other. This is not true for the selected dataset of the UK Covid-19 affected people of BBC UK news dataset.

Accessing the selected dataset it was observed that the selection of the alternative hypothesis helps to develop this data analysis study. To develop the statistical hypothesis the 'p' value significance level is used here. The 'p' value helps to identify the evidence of rejecting the Null hypothesis development concept for a specific dataset. The hypothesis significance level value (p) has a range between 0 to 1 values. Small values of 'p' or the level of significance support the strong rejection of the Null hypothesis (Kachiashvili, 2019, p. 123). Application of the 'p' value 0.10 indicates the borderline of developing statistical hypotheses based on the Null hypothesis development approach. On the other hand, the application of the 'p' value 0.01 indicates rejection of the Null hypothesis concept. In the case of using the 'p' value as 0.01, it indicates a 1% probability of supporting the Null hypothesis for the selected statistical evidence. To develop the test, another hypothesis significant value is used, where the 'p' value is equal to 0.02.

This condition also rejected the Null hypothesis by representing a 2% probability of supporting the Null hypothesis for the selected statistical evidence. Observation of all the levels of significance value helps to obtain that for the infection rate statistical analysis alternative hypothesis is the right choice. On that concern, it can be said that the infection rate of Covid-19 is different for the white people and BAME people of the country UK according to the selected statistical evidence.

Selection and justification of analyzing techniques

To develop the statistical analysis for analyzing the ratio of infection rate and the truthfulness of the findings two different models of statistics are used here. This data analysis section uses a t-test and chi-square test for the statistical data analysis purpose. For the data demonstration purpose and statistical method application phase, the SPSS software is used here. The application of the SPSS statistical software package helps to arrange the statistical data and support the hypothesis testing process. After the development of the hypothesis, the application of the t-test helps to identify the statistical significance of the developed hypothesis. Also, this data analysis and visualization concept uses the chi-square statistical test. The application of the chi-square statistical test helps to identify the relationship between the different variables of the selected dataset. To provide a logical data analysis and data visualization description both the t-test and chi-square test statistical test techniques are used in the SPSS software platform.

The Chi-square test provides the likelihood ratio for analyzing the relational probability of different variables of the dataset (Kim, 2019, p. 4257). The application of a statistical method or technique t-test is used here to obtain the statistical significance of the developed hypothesis for the Covid-19 infection trend for white people and BAME people of the country UK. The t-test considered is mainly three different factors of the selected statistical data such as; distribution of t, degree of freedom, and t statistics to demonstrate the statistical significance of the selected BBC UK news statistical data. The application of t-test helps to compare the age of infection occurrence for the two main columns of the selected data table such as; 'White' and 'BAME'. At the initial condition of the t-test, the application assumes the null hypothesis for the problem statement analysis.

Under the process of t-test, different values are calculated and compared with the standard value. This comparison process helps to identify the real statistical hypothesis for the selected dataset by rejecting or accepting the null hypothesis of the initial condition (Hamada, 2018, p. 18). To use the t-test technique three different values of the selected dataset have to be calculated in the initial condition such as; total numbers of data value, standard deviation value of the data column, and mean value of each data column. Results of the t-test help to obtain the right hypothesis for the dataset. On the other hand, the Chi-square test helps to identify the relationship between different variables. The application of this method helps to evaluate the cross-tabulation to identify the independence of the developed hypothesis. The cross-tabulation

of the SPSS software package helps to conduct the chi-square test. The application of the Chi-square test is quite justified for this data analysis section due to its approach presenting data relationships for identifying the truthfulness of the findings.

Selection and justification of techniques to present results

The statistical analysis method is used for presenting the results of the t-test and chi-square test. All of the input values and output values of the t-test and chi-square test having an integer value. In that context, the application of statistical analysis techniques is quite justified for this data result analysis purpose. Application of the t-test helps to outline the hypothesis test of the selected data sample. To analyze the results 'one sample test' method is used for the 't-test' statistical process. On that approach using the SPSS software technology is selected here for presenting the statistical results (Williams *et al.* 2017, p. 330). The SPSS technology helps to compare the sample mean value to the hypothesized population means to determine their difference. The application of the crosstab in SPSS technology helps to present the results of the chi-square statistical test. To present the chi-square test results and data analysis the SPSS technology is used here due to its features of contingency table statistical data analysis using cross-tabulation process.

Application

Application of the techniques analysis

Two different statistical methods are used for the analysis purpose by following the SPSS techniques of statistical results representation. This section helps to detail evaluate the application process of statistical data analysis of the selected BBC UK news data of Covid-19 infection rate for white people and BAME people of UK.

SPSS t-test technique application analysis

To test the hypothesis t-test is used here and the data presentation process uses the SPSS technique. For this t-test, the first part is to develop the 'one sample statistics' table in the SPSS platform. This one sample statistics table helps to identify the mean value, standard deviation value, error of the standard deviation value, and total data value (N). The application of all of these values helps to develop the data analysis t-test.

	N	Mean	Standard deviation	Standard error mean
White patient	50	36.18	9.419	1.332
BAME patient	50	36.74	9.317	1.318

Table1. Basic statistical calculation

(Source: Self-created)

After identification of those data, the SPSS technique is used for result presentation using t-test. For that purpose, SPSS t-test development steps are to click on Analyze > Compare Means > Independent-Samples T-Test (Narwaria *et al.* 2018, p. 2066).

SPSS chi-test technique application analysis

The application of the chi-square test using the SPSS software package helps to obtain the relationship of different data variables. To apply the chi-square test the first step is to develop the case summary using the feature of crosstab in SPSS. Under SPSS Clicking on Analyze -> Descriptive Statistics -> Crosstabs helps to create the crosstab for the chi-square test (Tuan *et al.* 2020, p. 890).

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
White patient*	50	100.0%	0	0.0%	50	100.0%
BAME						

Table2. Crosstab: case processing summary

(Source: Self-created)

The above table helps to obtain the crosstab of the Chi-square test. This observation also helps to analyze that in the selected dataset there are no outliers or missing values that can affect the

results. After the development of the crosstab in the SPSS software platform, the 'Statistics' option has to be selected. After that selection of the 'Chi-square' option helps to present the result data in the SPSS software.

Interpretation of the result analysis

Results of the t-test under the SPSS result presentation techniques

	Test value=0					
	t	df	Sig. (2 tailed)	Mean differenc e	95% Confidence Interval of the Difference	
					Lower	Upper
White patient	27.161	49	.000	36.180	33.50	38.86
BAME	27.883	49	.000	36.740	34.09	39.39

Table3. T-test results

(Source: Self-created)

Results of the chi-square test under the SPSS result presentation techniques

	Value	df	Asymp. Sig. (2- sided)
Pearson chi-square	544.692 ^a	484	.029
Likelihood ratio	206.914	484	1.000
Linear by linear	.252	1	.616

association	50		
N of Valid cases			

a. 529 cells (100.0%) have expected counts less than 5. The minimum expected count is .02.

Table4. Chi-square test result

(Source: Self-created)

Presentation of results

Analyzing the data of the t-test result table it was found that the mean difference for both cases of white people and BAME people the mean difference value is not equal to 0. In that context, it can be said that the hypothesis is not null. This observation helps to states the alternative hypothesis is significant for the dataset of BBC UK news of Covid-19 infection rate. The value of Pearson chi-square from Chi-test it was found that the real data distribution is different from the theoretical data distribution. The statistical value of the likelihood ratio helps to obtain the hypothesis value relation. This observation helps to identify the different trends of infection rates for white and BAME people in the UK.

Conclusion

Concluding this analysis it was observed that the application of the statistical data analysis helps to obtain the hypothesis testing results for the selected dataset. This analysis used the application of the t-test to help to test the three different significant levels of the hypothesis. The combined application of the t-test and Chi-test in the data presentation platform of SPSS helps to obtain the infection ratio and truthfulness of the covid-19 affected people results for the white and BAME people of the country UK. Both of the statistical analyses help to obtain that alternative hypothesis is suitable for this statistical dataset.

Reference list

- Hamada, C., 2018. Statistical analysis for toxicity studies. *Journal of toxicologic pathology*, 31(1), pp.15-22. https://www.jstage.jst.go.jp/article/tox/31/1/31_2017-0050/_article/-char/ja/
- Kachiashvili, K.J., 2019. Modern state of statistical hypotheses testing and perspectives of its development. *American Statistician*, 59(2), pp.121-126.
- Kim, P.S., 2019. Selective finite memory structure filtering using the chi-square test statistic for temporarily uncertain systems. *Applied Sciences*, 9(20), p.4257.
- Narwaria, M., Krasula, L. and Le Callet, P., 2018. Data analysis in multimedia quality assessment: Revisiting the statistical tests. *IEEE Transactions on Multimedia*, 20(8), pp.2063-2072.
- Tuan, B.A., Pudikova, G.N., Mahmoudi, M.R. and Pho, K.H., 2020. Statistical approaches in literature: Comparing and clustering the alternatives of love in Divan of Hafiz. *Digital Scholarship in the Humanities*, 35(4), pp.886-892.
- Wendler, T. and Gröttrup, S., 2016. *Data mining with SPSS modeler: theory, exercises and solutions*. Springer.
- Williams, M.N., Bååth, R.A. and Philipp, M.C., 2017. Using Bayes factors to test hypotheses in developmental research. *Research in Human Development*, 14(4), pp.321-337.
- Yusup, M., Naufal, R.S. and Hardini, M., 2018. Management of Utilizing Data Analysis and Hypothesis Testing in Improving the Quality of Research Reports. *Aptisi Transactions on Management (ATM)*, 2(2), pp.159-167.

