Epidemic Spread Pattern Detection Model

Ipshita Singh
MKSSS's Cummins College of Engineering for Women, Pune, India

Abstract

The onset of an epidemic leads to a chaotic situation as there is no understanding of the pattern of spread of the disease. Hence an effective use of combination of statistical tests and visualized data can help predict patterns and trends in medical data which can save valuable time. Statistical values can help understand dependence of factors on occurrence of the disease. Whereas visual reports based on age, gender, co-morbid conditions of disease can help understand pattern of spread and also help in early control of an epidemic outbreak. If these reports could be monitored regularly during the outbreak medical aid, funding and precautionary measures can be taken accordingly focusing on adversely affected areas. A combined model of rightly chosen statistical tests and a combination of visualization reports is developed using R programming and Tableau respectively. Results and effectiveness of tests have been discussed based on implementation done on swine flu data of regions of Maharashtra, India during 2009 epidemic outbreak. Implementation and study has been done using chi square test and visuals generated by Tableau.

Keywords

Data Visualization, Statistical Analysis, Swine Flu, Epidemic, R Programming, Tableau, Chi Square Test

I. Introduction

"Health is wealth" is a phrase which is true to its words. In situations of epidemics not only is it a loss of lives but it also takes a huge hit on the country's economy. Hence it is the need of the hour to work towards understanding the epidemic nature and its spread. An epidemic can be stated as a situation of rapid spread of an infectious disease to the masses in a given population within a short period of time. A combination of statistical analysis and data visualization can provide us with a deeper insight to the same.

A. Statistical Analysis in Medicine

Statistical analysis is a powerful tool to draw meaningful conclusions from data which is collected through survey, observation or experimentation. Depending on the data set and the kind of conclusions required appropriate statistical test can be chosen. The outcomes of these results are interpreted along with indicating the level of uncertainty involved. Following are a few methods that are being used widely in the field of medicine:

1. ANOVA

ANOVA stands for Analysis of Variance. It provides statistical tests to check if the mean of several groups are equal or has differences. It is checked within the groups and also among the groups.

2. Chi-Square Test

Chi-Square Test is used to test hypothesis of relation between two groups or populations.

3. Logistic Regression

Logistic regression analysis is a type of regression analysis used for finding risk factor and predicting the outcome of a categorical criterion variable based on one or more predictor variables. Either one of these tests or a combination of two or more can be used create models for analyzing medical data.

B. Data Visualization

Data Visualization can be defined as an effective means of communicating qualitative data to viewers through graphs, plots, tables and charts. It is necessary for understanding patterns, trends and relationships that exits in the data.

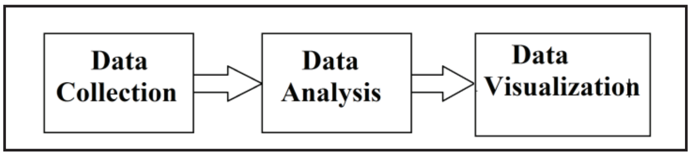

Fig. 1: Approach to Data Visualization

A wide variety of visualization tools are present these days which can help understanding our data effectively.

II. Chi Square Test

A. Formula

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

Where,
O_i = observed frequency in the ith cell of the table
E_i = expected frequency in the ith cell of the table
R_t= row total corresponding to I cell
C_t= column total corresponding to I cell
E_i= (R_t*C_t)/Total

1. Formula for Degree of Freedom (df):

df=(r-1)*(c-1)
Where,
r = number of rows in the table
c = number of columns in the table

B. Probability Value or p-value

These values are found using chi-square statistics distribution table. The row along the degree of freedom value is chosen. Within that row a value greater than the calculated chi-square value is chosen. The corresponding column of this chosen value and the next greatest value determines the p-value range.

If p <= significance level statistically significant.
If p > significance level not statistically significant.

C. Significance Level

The significance level is used to determine whether the null hypothesis can be rejected or accepted by comparing it with the p-value. By convention 0.05 or 5% is the chosen value.

D. Interpretation of the test

If p-value is less than significance level then the null hypothesis can be rejected stating that a relationship exists between the variables. If the value is greater than 0.05 then we can say that it is not statistically significant and that there is high chance that the variables are independent.

III. Tableau

Tableau is Business intelligence software that allows the user to connect to multi-platform data easily and create interactive visuals within split seconds. It is easy enough for even an excel user to understand and powerful enough for solving extremely complex analytical queries. It queries spreadsheets, cubes, cloud databases and relational databases to generate required number of graphs that can be combined into dashboards or interactive stories and shared over a computer network or the internet.

A. Highlighted Features of Tableau

1. User friendly and easy installation.
2. Easy to make dashboards which help understanding data and simultaneous parameters affecting it in a short span of time. These dashboards can then be shared with everyone necessary on Tableau server or online (BI cloud).
3. Answers to millions of data in rows can be obtained in seconds.
4. Simple drag and drop functions, no need of complex wizards or scripting.
5. Tableau public and Tableau Reader versions are free to use.

IV. Implementation

Data available is from parts of Maharashtra namely sangli, Pune, Miraj, Latur. This data has been divided as follows for our combinational tests.

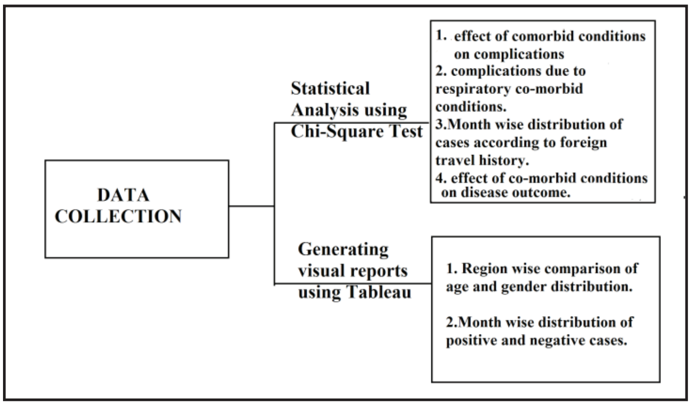
DATA COLLECTION
Statistical Analysis using Chi-Square Test
1. effect of comorbid conditions on complications
2. complications due to respiratory co-morbid conditions.
3.Month wise distribution of cases according to foreign travel history.
4. effect of co-morbid conditions on disease outcome.
Generating visual reports using Tableau
1. Region wise comparison of age and gender distribution.
2.Month wise distribution of positive and negative cases.

Fig. 2: Distribution of Data for Implementation

A. Chi Square Test

Table 1: Respiratory Co-morbid Conditions and Complications in Cases [1]

	Complications Present	Complications Absent
Comorbid Present	9	12
Comorbid Absent	62	201

On passing the data mentioned in the table to an inbuilt function chisq.test() in R, it returns the χ^2 value as 3.8566 and p-value as 0.04955. Since the p value is less than 0.05(significance level) we can reject the null hypothesis and state that presence of respiratory co-morbid conditions does effect the development of complications among patients. Along with data set we pass “correct=F” as a parameter. This is for ignoring the yate’s correction.

Table 2: Month Wise Distribution of Cases According to History of Contact During the 8 Days Prior to Onset of Symptoms [1]

	ForeignTravel Present	ForeignTravel Absent
June	3	0
July	11	68
August	12	127
September	0	53
October	0	10

The p-value obtained here is 9.607e-08 which is very less than 0.05 and hence we can state the outcome is highly statistically significant. There is definitely a relation between the occurrences of the disease in a patient to the month in which he travelled to an affected foreign country.

Table 3: Association of co-morbid Condition and Mortality Among Positive Cases [3]

	Comorbid Present	Comorbid Absent
Survived	13	85
Deaths	8	2

The p-value obtained here is 3.787e-07 which is less than 0.05 hence it is highly significant. This proves the existence of relation between survival and deaths with presence and absence of co-morbid conditions.

Table 4: Effect of Non Respiratory Comorbid Conditions on Complications [1]

	Complications Present	Complications Absent
Comorbid Present	15	30
Comorbid Absent	56	183

The p-value obtained is 0.1593. Since p-value is greater than 0.05 we can there is relation of occurrence of complications in patients due to due an existing non respiratory co morbid condition.

B. Tableau

Table 5: Age and Gender Distribution in Pune and Sangli-Miraj [1-2]

age	female	Male	region
0 to 20	109	86	pune
21 to 40	24	47	pune
41 to 60	11	4	pune
>60	1	2	pune
0 to 20	10	24	Sangli - miraj
21 to40	32	35	Sangli - miraj
41 to 60	15	17	Sangli - miraj
>60	4	5	Sangli - miraj

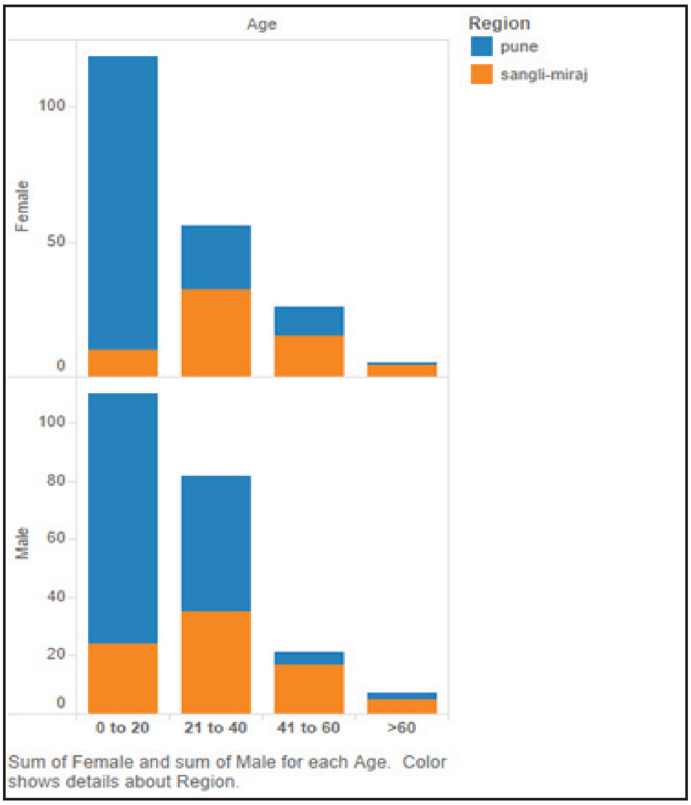

Fig. 3: Region Comparison of Male and Female Gender Wise Distribution Created Using Tableau

Table 6: Month wise distribution of positive and negative cases from October 09-September 10 in sangli and miraj region [2].

Month	Positive	Negative
2009 October	18	34
2009 November	0	5
2009 December	6	7
2010 January	3	6
2010 February	4	8
2010 March	2	5
2010 April	1	3
2010 May	0	3
2010 June	0	2
2010 July	18	23
2010 August	73	156
2010 September	17	72

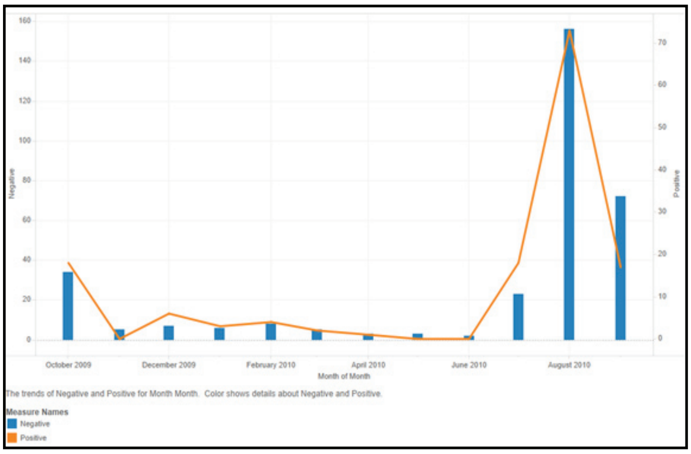

Fig. 4: Month Wise Distribution of Cases

The objective of visualization is pretty clear from these examples. Within seconds of looking at the images we could make many conclusions regarding most affected gender, age group, peak month of positive cases etc.

Tableau also offers the feature of creating highly interactive Dashboards. These Dashboards are a combination of graphs such as above. They help in monitoring various situations in one single view. This can be an additional advantage while including this as a actual usable application at medical centers.

V. Results

The results obtained from the above conducted chi square tests tell us that a history of foreign travel made the patient more prone to the disease. The presence of respiratory co-morbid conditions leads to a rise of complications in patient whereas presence of non respiratory co morbid conditions had no such effect. Finally the presence of a co-morbid condition leads to a higher mortality. Similarly in case of visuals obtained we can conclude from the male female gender wise distribution visual that in both regions the age group between 0-20 registered maximum cases with larger number of males being affected in sangli-miraj as compared to males in pune and vice versa for female. Using the month wise distribution visual we can conclude that august 2010 registered highest number of negative as well as positive cases in sangli-miraj region.

VI. Conclusion

A. Summary

This report has aimed at understanding the pattern of spread through statistical analysis test and data visualization. The chi square test gives us the existence of relationship between variables which can be a guiding factor in making decisions for elimination of the spread of swine flu. Visuals generated using Tableau can help predict the area's most affected and dependence on factors.

B. Future enhancements

Statistical Tests like ANOVA and logical regression can also be included in this model. This will help in further narrowing down the relationships between variables and groups of values. A tool or application can be generated which can be a combination of these statistical tests and highly representative dashboards. This tool can then be deployed at medical centers for monitoring the situation of any kind of epidemic outbreak and help them take the right steps towards controlling the disease spread and providing sufficient medical aid to necessary people.

References

[1] M.P Tambe, M Parande, A.V Jamkar, R.R Pardesi, K Baliwant, P.S Rathod, S.D Patsute, A.P Todsam RT Pote,"An Epidemiological Study of Confirmed H1N1 Admitted Cases in an Infectious Disease Hospital", Pune: JKIMSU, Vol. 1, No. 2, July-Dec. 2012

[2] J. D. Naik, Kriti A. Patel, S.S. Rajderkar, Kailas R Bhoye," H1N1 Swine flu: An Experience in a district of Western Maharashtra", India: International Journal of Collaborative Research on Internal Medicine & Public Health

[3] Chinte L. T., V. V. Kendre, Godale L. B,"Study of Clinico - Epidemological Features of the Hospitalized Patients of Confirmed Influenza A (H1N1) Virus Infection in Government Medical College", Latur, Maharashtra: International Journal of Recent Trends in Science And Technology, Vol. 10, Issue

2, 2014 pp. 399-402

[4] Chan-Yong Park, Joon-Ho Lim, Han. Hyung Soo, “Statistical Analysis Service of e-Healthcare Record on iPad System”, South Korea: 2014 IEEE International Conference on Consumer Electronics (ICCE)

[5] [Online] Available: https://en.wikipedia.org/wiki/2009_flu_pandemic

[6] [Online] Available: http://www.slideshare.net/parth241989/chi-square-test

[7] [Online] Available: http://www.tableau.com/products/technology

[8] [Online] Available: http://www.flu.gov/about_the_flu/h1n1/