# CHAPTER IV

# METHODOLOGY

## **4.1 Data Collection and Processing**

## **4.2 Statistical Methods**

For simultaneous and systematic interpretation, multivariate statistical techniques were used to interpret the antibiotic data and to give meaningful results that were not possible while assessing the data at a glance. The data have been analyzed by using RStudio software.

### **4.2.1 Descriptive Statistics**

The descriptive statistics such as the Mean, Minimum value, Maximum value, Standard Deviation (SD) and Coefficient of Variation (CV) were computed for each variable. Descriptive statistics gives the summary about the huge data. Through this we can observe the variation in variables values.

### **4.2.3 Correlation Coefficient Test**

A correlation coefficient test were used to assess the degree of dependency of one cluster to another or how strong the relation is between two variables.

### **4.2.4 Pearson’s Correlation Coefficient**

In this study, Pearson’s correlation analysis was utilized to calculate the correlation between two variables in order to characterize the relationship of the all variables. For a correlation between variables x and y, the formula for calculating the sample Pearson’s Correlation Coefficient is given by,

r =

*where xi and yi are the values of x and y for the ith individual.*

### **4.2.5 Partial Correlation**

Partial correlation is the measure of association between two variables, while controlling or adjusting the effect of one or more additional variables. Partial correlations can be used in many cases that assess for relationship, like whether or not the sale value of a particular commodity is related to the expenditure on advertising when the effect of price is controlled. Formula for partial correlation coefficient for X and Y with controlling for Z

### **4.2.6 Data Standardization**

Data standardization is essential in multivariate analysis; it increases the homogeneity of a dataset and enhances its normality, and therefore ensures that all parameters are close in terms of their variances (Sandow Mark Yidana et al., 2010). The data were therefore standardized to their corresponding z-scores as in Equation below, in order to achieve the objectives of normal distribution and homogeneity.

Where, *x is the data, and s is the standard deviation of the datasets.*

## **4.3 Identify Different Factors**

An attempt was made to identify the factors among all variables which are contributing more to the total variability in whole dataset.

### **4.3.1 Factor Analysis**

Factor analysis (FA) is a multivariate statistical technique, which gives the general relationship between measured variables by showing multivariate patterns that may help to classify the original data. FA is designed to transform the original variables into new uncorrelated variables called factors, which are linear combinations of the original variables. It is a dimension reduction technique and suggests how many variables are important to explain the observed variances in the data.

Factor analysis was also applied to the data to determine and rank the sources of variation in the antibiotic data. In the factor analysis, ‘principal components’ method was selected as the solution method. The total number of factors generated from a typical factor analysis indicates the total number of possible sources of variation in the data. Factors are ranked in order of merit. The first factor or component has the highest eigen value and represents the most important source of variation in the data. The last factor is the least important process contributing to the variation. Factor loadings are interpreted as correlation coefficients between the variables and the factors.

This treatment provides a small number of factors that usually account for approximately the same amount of information as the original set of observations. The observed variables are modeled as linear combinations of the potential factors including the error terms as follows:

*Zij* = *a*1*f*1*j* + *a*2 *f*2*j* +....+ *amfmj* + *eij*

*Where  
Zij= Measured variable, ai= ith Factor loading, fij= Factor score, eij= Error term  
i=1, 2, 3,…m, j=1, 2, 3,…p*

### **4.3.2 Cluster Analysis**

Cluster analysis is a multivariate technique first used by Tryon, (1958) which attempts to combine cases into groups when the group membership is not known prior to the analysis. That is, cluster analysis is a technique for grouping individuals or objects into unknown groups. These groups are relatively homogeneous within themselves and heterogeneous between each other, on the basis of a defined set of variables. These groups are called clusters.

In biology, cluster analysis has been used for decades in the area of taxonomy, where living things are classified into arbitrary groups on the basis of their characteristic groups. The classification proceeds from the most general to the most specific in steps. The most general classification was kingdom followed by phylum, subphylum, and class, etc.

### **4.3.2.1 Clustering Methods**

The commonly used methods of clustering fall into two general categories. (i) Hierarchical (ii) Non hierarchical. Hierarchical clustering is the most commonly used method of clustering; which proceed by either a series of mergers or a series of successive divisions. In the initial stage no need to take decision regarding the number of clusters.

The main objective of this classification approach is to determine the relationships and associated chemical processes between variables. In HCA classification, data log-transformation and standardization is required for equal weighting because parameters with higher or smaller variance during their distribution affects the Euclidean distance calculation (Lu et al., 2015).

### **4.3.2.3 Euclidean Distance**

Euclidean distance is the most commonly used measure of distance in two dimensions, one for the plot of observations in a scatter diagram, and another to measure the distances between the pairs of points. Generally, the following equation can be used as a distance measure:

d(x,y) =Σi (xi−yi)2

It divides datasets into hierarchies based on similarity or dissimilarities in the field. In this study the Qmode HCA, groups pumps locations into clusters. The method is used for clustering is Squared Euclidean Distance which is the distance between two items, x and y, is the sum of the squared differences between the values for the items.

### **4.2.3.4 Dendrogram Presentation**

The Dendrogram is a graphical representation of the results of hierarchical cluster analysis. This appears in the form of a tree like plot where, each step of hierarchical clustering is represented as a fusion of two branches of the tree resolving into a single one. The branches represent clusters obtained on each step of hierarchical clustering.

The HCA resulted in a dendrogram which is a presentation of the antibiotic associations in the area. Samples with similar characteristics and relationships are clustered together at low linkage distances, whilst dissimilar samples are linked at higher linkage distances.

# CHAPTER VI

# RESULTS

## **6.1 Descriptive Statistics**

Examination of the statistical characteristics of the antibiotic analysis shows that the AMP values of antibiotic ranged from 0 to 19 with an average value of 0.88. PIT value ranged from 0 to 46 with a mean of 30.04. The value of CTX varied from 0 to 35 with an average value of 15.46. The concentration of CTR is observed from 0 to 36 with the mean value of 18.08. GEN varied from a low of 0 to a high of 30 with the mean value of 15.79. CX ranged from 0 to 30 with the mean value of 4.28. AZM value varied from 0 to 32 with the mean value of 13.65. TE varied from 0 to 40 with the mean value of 15.94. The COT, CPM, IMP, LE and CXM varied between 0 to 32, 0 to 34, 0 to 42, 0 to 40, and 0 to 24 with the mean of 11.39, 17.29, 27.88, 24.07, and 2.12, respectively.

**Table 6.1: Degree of Dispersion and Standard Specifications of antibiotic Concentrations for patients in ………..\***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameters | Mean | SD | Min | Max | CV |
| AMP | 0.88 | 3.72 | 0.00 | 19.00 | 4.25 |
| PIT | 30.04 | 6.84 | 0.00 | 46.00 | 0.23 |
| CTX | 15.46 | 10.75 | 0.00 | 35.00 | 0.70 |
| CTR | 18.08 | 11.28 | 0.00 | 36.00 | 0.62 |
| GEN | 15.79 | 8.11 | 0.00 | 30.00 | 0.51 |
| CX | 4.28 | 8.69 | 0.00 | 30.00 | 2.03 |
| AZM | 13.65 | 9.29 | 0.00 | 32.00 | 0.68 |
| TE | 15.94 | 9.88 | 0.00 | 40.00 | 0.62 |
| COT | 11.39 | 11.93 | 0.00 | 32.00 | 1.05 |
| CPM | 17.29 | 12.59 | 0.00 | 34.00 | 0.78 |
| IMP | 27.88 | 8.87 | 0.00 | 42.00 | 0.32 |
| LE | 24.07 | 8.06 | 0.00 | 40.00 | 0.33 |
| CXM | 2.12 | 5.72 | 0.00 | 24.00 | 2.70 |

**\*SD = Standard Deviation, CV = Coefficient of variation**

Table 6.1 also depicts that the standard deviation of some of the parameters are higher than the average value (in terms of CV), and it indicates that the antibiotic of the study area is not homogenous. The diversity of antibiotic data indicates that AMP, CX, COT, and CXM content were highly varied (CV>1), then CTX, CTR, GEN, AZM, TE, and CPM were moderately varried (0.50 < CV < 1) and the rest were low varried (CV<0.50). Central value of each parameter alone with scatterness are presented in the Figure 6.2 for the comprehension at a glance.



## 6.3 Correlation Analysis

As given in the Table 6.3, out of 78 pairs of antibiotic attributes, 44 pairs were found significantly correlated (p<0.05) and the frequency of correlation indicated that the patient characteristics could be grouped into factors based on their correlation patterns. There is very strong positive significant (p <0.05) correlation at between CPM and CTR (r = 0.76), moderate positive correlation between IMP and PIT (r = 0.52), CPM and CTX (r = 0.53).

**Table 6.3: Correlation Coefficient among antibiotic variables**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | AMP | PIT | CTX | CTR | GEN | CX | AZM | TE | COT | CPM | IMP | LE | CXM |
| AMP | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |
| PIT | **0.20** | 1.00 |  |  |  |  |  |  |  |  |  |  |  |
| CTX | **0.39** | **0.39** | 1.00 |  |  |  |  |  |  |  |  |  |  |
| CTR | **0.38** | **0.29** | **0.61** | 1.00 |  |  |  |  |  |  |  |  |  |
| GEN | 0.07 | **0.25** | 0.02 | **0.28** | 1.00 |  |  |  |  |  |  |  |  |
| CX | **0.53** | **0.20** | **0.38** | **0.35** | 0.05 | 1.00 |  |  |  |  |  |  |  |
| AZM | **0.26** | **0.33** | **0.19** | **0.29** | **0.45** | 0.14 | 1.00 |  |  |  |  |  |  |
| TE | **0.32** | **0.20** | -0.11 | -0.02 | **0.47** | **0.20** | **0.31** | 1.00 |  |  |  |  |  |
| COT | **-0.23** | -0.11 | **-0.18** | -0.15 | -0.02 | -0.07 | -0.13 | 0.16 | 1.00 |  |  |  |  |
| CPM | **0.37** | 0.12 | **0.53** | **0.76** | **0.21** | **0.23** | **0.41** | -0.07 | **-0.19** | 1.00 |  |  |  |
| IMP | **0.23** | **0.52** | **0.24** | **0.22** | 0.14 | **0.21** | **0.22** | 0.10 | **-0.27** | 0.13 | 1.00 |  |  |
| LE | **0.28** | 0.09 | 0.03 | **0.17** | 0.15 | 0.16 | 0.10 | **0.42** | **0.21** | 0.11 | -0.01 | 1.00 |  |
| CXM | -0.09 | 0.15 | 0.11 | **0.42** | 0.08 | **0.20** | -0.11 | 0.04 | 0.17 | 0.11 | 0.01 | 0.03 | 1.00 |

**\*Correlation is significant at the 0.05 level (two-tailed)**

## **6.4 Partial Correlation Analysis**

A partial correlation coefficient enumerates the correlation between two or more variables when removing the effects of one or several other variables (de la Fuente, Bing, Hoeschele, & Mendes, 2004). The partial correlation coefficient matrix for the antibiotic variables is given in Table 5.4. The partial correlation coefficients among 13 antibiotic variables were calculated for correlation analysis. When removing the effects of all other variables, AMP shows significant negative significant partial correlation with GEN (-0.25), COT (-0.21) and CXM (-0.30), and significant positive correlation with CTR (0.20), CX (0.41), and TE (0.38).

**Table 6.4: Spearman Partial Correlation Coefficient among antibiotic variables**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | AMP | PIT | CTX | CTR | GEN | CX | AZM | TE | COT | CPM | IMP | LE | CXM |
| AMP | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |
| PIT | 0.01 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |
| CTX | 0.08 | **0.36** | 1.00 |  |  |  |  |  |  |  |  |  |  |
| CTR | **0.20** | 0.02 | **0.30** | 1.00 |  |  |  |  |  |  |  |  |  |
| GEN | **-0.25** | 0.09 | -0.13 | **0.30** | 1.00 |  |  |  |  |  |  |  |  |
| CX | **0.41** | -0.08 | **0.21** | 0.01 | -0.02 | 1.00 |  |  |  |  |  |  |  |
| AZM | -0.01 | **0.26** | -0.10 | -0.01 | **0.23** | 0.05 | 1.00 |  |  |  |  |  |  |
| TE | **0.38** | 0.05 | -0.09 | **-0.23** | **0.47** | 0.05 | 0.17 | 1.00 |  |  |  |  |  |
| COT | **-0.21** | 0.01 | 0.04 | -0.05 | -0.06 | 0.04 | -0.02 | 0.16 | 1.00 |  |  |  |  |
| CPM | 0.08 | **-0.23** | 0.16 | **0.58** | -0.01 | -0.08 | **0.33** | -0.09 | -0.02 | 1.00 |  |  |  |
| IMP | 0.02 | **0.44** | -0.07 | 0.10 | -0.03 | 0.09 | -0.01 | 0.04 | -0.19 | -0.03 | 1.00 |  |  |
| LE | 0.10 | 0.07 | -0.11 | 0.19 | -0.09 | 0.01 | -0.08 | **0.33** | **0.21** | 0.01 | -0.08 | 1.00 |  |
| CXM | **-0.30** | 0.16 | -0.19 | **0.54** | -0.11 | **0.23** | **-0.21** | 0.19 | 0.14 | -0.15 | -0.11 | -0.14 | 1.00 |

**\*Correlation is significant at the 0.05 level (two-tailed)**

## **6.5 Component Matrix with Total Variance Explained**

Principal component analysis (PCA) was performed on standardized data sets (13 variables × 112 patients) to reduce the dimensions of the original data sets and to identify latent factors affecting antibiotic data. The number of significant principal components (PCs) is determined based on both scree plots and eigenvalue. Eigenvalue that indicates that PCs with eigenvalues are considered important when the correlation matrix is used in the analysis (Cattell, 1966). Factor scores represent the cumulate contribution of all parameters loaded on a particular factor/principal component. In this study, as presented in Table 6.5, PCA extracted five significant PCs with eigenvalues > 1, explaining about 74.09% of the total variance in corresponding antibiotic data sets of the study area. The examination of the Scree plot as shown in Figure 6.5 provides a visualization of the variance associated with each factor, the steep slope shows the biggest factors. It is clear from the figure that there are five dominant factors of the total variance of the antibiotic.

**Table 6.5: Rotated factor loadings for antibiotic samples in the study area (Extraction method\*\*)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Parameters** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** |
| AMP | **0.312** | -0.119 | 0.132 | **-0.524** | 0.249 |
| PIT | 0.234 | 0.126 | **0.421** | 0.070 | **-0.495** |
| CTX | **0.384** | 0.198 | -0.077 | -0.090 | -0.016 |
| CTR | **0.417** | 0.118 | **-0.350** | 0.114 | 0.013 |
| GEN | 0.283 | -0.286 | 0.017 | **0.438** | 0.037 |
| CX | **0.310** | -0.101 | -0.024 | **-0.532** | -0.060 |
| AZM | 0.288 | -0.177 | 0.266 | **0.408** | 0.183 |
| TE | 0.079 | **-0.574** | 0.283 | -0.030 | -0.122 |
| COT | -0.156 | **-0.379** | -0.259 | 0.055 | **-0.309** |
| CPM | **0.366** | 0.073 | **-0.328** | 0.202 | 0.293 |
| IMP | 0.271 | 0.269 | **0.323** | 0.041 | **-0.327** |
| LE | 0.132 | **-0.492** | -0.117 | -0.111 | 0.011 |
| CXM | 0.109 | -0.026 | **-0.482** | -0.013 | **-0.592** |
| **Eigenvalue** | 3.70 | 1.92 | 1.44 | 1.36 | 1.22 |
| **Variance Explained (%)** | 28.48 | 14.75 | 11.06 | 10.45 | 9.35 |
| **Cummulative Variance (%)** | 28.48 | 43.23 | 54.29 | 64.74 | 74.09 |

\*\* Five components extracted from matrix; Extraction method = Principal component analysis; Loadings greater than 0.3 are in bold.

The rules of classifying the factor loadings as "strong," "medium," and "weak," corresponding to absolute loading values of >0.75, 0.75– 0.50, and 0.50–0.30, respectively (Liu, Lin, & Kuo, 2003). It clearly shows that first few factors explain a relatively larger amount of variance. Whereas, the subsequent factors explain the small amount of variance. Those factors variance exceeded 70% and they are sufficient to explain the mechanisms that controlling whole datasets (Nagaraju, Sunil Kumar, Thejaswi, & Sharifi, 2014).

|  |
| --- |
|  |
| **Figure 6.5: Scree plot with % of explained variance** |

PC1, accounting for 28.5% of the total variance, had weak positive loadings on AMP, CTX, CTR, CX, and CPM. PC2, accounting for 14.75% of the total variance, has moderate negative loadings on TE. PC3, accounting for 11.06% of the total variation, has weak positive loadings on PIT and IMP. In addition, PC4 accounts for about 10.45% and it is moderately negative loaded with AMP (-0.524) and CX (-0.532). PC5 accounts for about 9.35% and it is moderately negative loaded with CXM (-0.592).

Many methods have been developed to provide multivariate data insights using interactive visualization, scatterplot, Parallel Coordination Plot (PCP) and Radar Chart as important representatives. On the radar chart (many other names, such as the Spider Chart, Star Chart, and Kiviat diagram) is a data visualization system that describes the multi-dimensional data on the two-dimensional plane (Claessen & van Wijk, 2011). It depicts and allows the comparison of several variables over time (Stafoggia et al., 2011). Rotated PCA was considered because of the logical grouping of variables under different factors and their high frequency of correlation with individual characteristics. Radar charts highlight those parameters that are undergoing the most or the least effect on the contribute to PCs. With the use of overlying data, changes in any parameters are visually depicted (Figure 6.6). Radar (spider) chart illustrates the loadings of PCA of antibiotic parameters of five dominant PCA. Data for the current PC1 (orange) are overlaid on data from the four previous factors (green = PC2, red = PC3, purple = PC4, and grey = PC5). Data represent (anti-clockwise from right) the loadings of the parameter.

|  |
| --- |
|  |
| **Figure 6.6: Rotated Factor Loadings of Five antibiotic Factors Extracted by PCA** |

## **6.6 Cluster Analysis**

The affiliation among the antibiotics was obtained through cluster analysis. Cluster analysis (CA) was employed to identify groups of similar characteristics. As the distance between the pair of antibiotics increases, it indicates less similarity between the antibiotics. The Euclidian distance between the antibiotics of the patients. It is one of the methods to measure the similarity between the antibiotics of the patients. By looking into the squared Euclidian distance table, it was observed great variability between the pumps (Appendix 6.2).

Essentially, data points with the smallest distances between them are grouped together. Then the data with the next smallest distances are added to each antibiotics, etc. until all observations end up together in one large group. The cluster is interpreted by observing the pattern produced. These may have some practical meaning in terms of the research problem. For this reason, hierarchical cluster analysis was used to detect the similarity groups between the sampling sites. Since hierarchical agglomerative cluster analysis was used, the number of clusters was also decided by the practicality of the results as there was information available on the study sites. It generated a dendrogram, grouping the 13 antibiotics into five distinct clusters (Figure 6.6).

|  |
| --- |
|  |
| **Figure 6.9: Dissimilatiy Matrix Based on Euclidean Distance** |

Results in Table 6.6 shows the history of the clustering process. In our case, CA has indicated five clusters or groups (Group E are the major cluster and Group A, B, and D are the minor clusters).

Table 6.6: Cluster Groups and their members

|  |  |  |
| --- | --- | --- |
| Group | Members (Location/Sample No.) | N (%) |
| A | PIT, IMP | 2 () |
| B | AMP, CX | 2 () |
| C | CTX, CTR, CPM | 3 () |
| D | COT, CXM | 2 () |
| E | GEN, AZM, TE, LE | 4 () |

## 6.7 Dendrogram HCA

|  |
| --- |
|  |

# References