**Introduction: -**

Diabetic Nephropathy (DN), or diabetic kidney disease, is a major complication of diabetes mellitus (DM). DM refers to a characteristic set of structural and functional kidney abnormalities in patients with diabetes. It is an irreversible and progressive disease and approximately 40% of people with type 2 diabetes develop nephropathy and it is a leading cause of end-stage renal disease (ESRD). DN can be quantitatively defined as a Glomerular filtration rate (GFR) <60 ml/min/1.73m2. In addition, rise in serum creatinine (SrCr) has been found to be strongly correlated with the presence of renal complication and is a predictor of DN in type 2 diabetic patients [4, 5, 6]. Studies have shown that poor glycaemic control (Fasting Blood Glucose (FBG)) and long duration of diabetes are the risk factors leading to progression of diabetic nephropathy with decline in GFR earlier

To build our detection model we will use Multivariate analysis approach the main reason to us the multivariate is that it automatically captures correlations between features while bin case of the original model we will have to manually create features for all different anomalies (problematic cases). One drawback of using multivariate is that it is computationally expensive. Response variables under consideration are often described as random variables and since their dependence is one of the things to be accounted for in the analyses, these response variables are often described by their joint probability distribution.

We have analysed the joint distribution of three correlated random variables: duration of disease, SrCr and FBG.  This is done by considering two data sets of type 2 DN patients namely dataset 1 and dataset 2. The first dataset gives the complete information (from the time of diagnosis of disease till the termination of study) of the renal health status of a type 2 diabetic patient and the second dataset gives the latest information about the health status of a patient collected through pathological reports of 19 months. The complete information of the three variables (duration of disease, SrCr and FBG) from dataset 1 is used to estimate the duration of disease for the patients belonging to the second dataset. Multivariate analysis is applied for estimating these disease durations by using the following procedure: firstly, appropriate distributions for three random variables namely duration of diabetes, SrCr and FBG from the first dataset are obtained by fitting distributions and appropriate distributions are selected on the basis of Akaike Information Criterion (AIC) for each random variable. Secondly, normal approximation is checked for each distribution. Lastly, we have checked the multivariate normality of duration of disease, SrCr and FBG by applying Mardia test. Now since the above three random variables are correlated, and are found to be

marginally and jointly normally distributed, it can be concluded that three dimensional MVN distribution is an appropriate joint distribution. Then the duration of disease is obtained by applying conditional expectation under MVN distribution for given values of SrCr and FBG.

This estimation procedure will help medical fraternity to guide those patients who have incomplete record history about their approximate duration of disease. Also, it will help in monitoring and evaluating the severity of DN complication. In fact, this paper will be more interesting for the statisticians to explore the applications of MVN and BVN distributions on real data.

**Workflow of the project: -**

Methodology: -

Here we will understand about all the different types of distributions and used in the process.

1. Multivariate Normal (MVN) Distribution

The random vector x is said to have p-dimensional multivariate normal distribution with a mean vector μ and variance-covariance matrix Σ if its joint probability density function is,

##### MULTIVARIATE NORMAL EXAMPLES: -

##### 

##### ; (similarly)

MULTIVARIATE NORMAL EXAMPLE

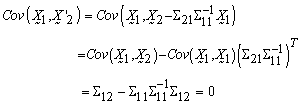
##### Here, μ is variance-covariance vector.

##### And sigma vector gets converted to;

##### Now we will define a transformation of (X1, X2) to new variables X1 and . This is achieved by using linear transformation,

##### 

##### As any linear combination of X is also MVN hence, the linear transformation shows that ( X1, X’2) are jointly MVN distributed. Now, we can show X1 and X’2 is independent by proving that they are uncorrelated.



##### Conditional Expectation under MVN Distribution

##### Let X be a p dimensional matrix. We shall divide this X into two

##### sub-vectors with dimension q and p-q. So, now;

##### ; (similarly)

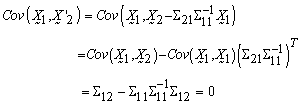
##### Here, μ is variance-covariance vector.

##### And sigma vector gets converted to;

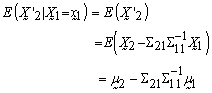
##### Now we will define a transformation of (X1, X2) to new variables X1 and . This is achieved by using linear transformation,

##### 

##### As any linear combination of X is also MVN hence, the linear transformation shows that ( X1, X’2) are jointly MVN distributed. Now, we can show X1 and X’2 is independent by proving that they are uncorrelated.



Since,  and  are MVN variables and uncorrelated they are independent. Thus,



Now, as  the conditional distribution of  given  is,

|  |
| --- |
|  |

Similarly, 

1. Bivariate Normal (BVN) Distribution

This distribution is a special case of multivariate Normal distribution with n=2. And its probability density is given by

and conditional expectation under BVN distribution is given as,

#### 3 Normal Approximation to Gamma and Lognormal Distributions

When the shape of the distribution curve becomes larger than we use Gamma distribution or Lognormal distribution so that the curve become better to evaluate and in a gaussian form. Gamma Distribution is given by:

But if the shape parameter γ is large as compared to λ then Gamma distribution tends to normal distribution i.e.,

In case of lognormal if arithmetic mean m is much larger than its arithmetic standard deviation s, then the distribution tends to Normal (m, s2). (say m > 6s). It is given by:

The mean and standard deviation can be defined as, &

#### Mardia Test for Multivariate Normality

Before applying any distribution to the data, we should check if it satisfy the underlying distributional assumptions. For most multivariate analyses, it is thus very important that the data indeed follow the multivariate normal or if not exactly at least approximately. The assumption is checked using univariate normality using PP plot or some other graphs. But Mardia multivariate skewness and kurtosis measures is a more convenient test to check our assumptions. The sample measures of multivariate skewness and kurtosis are,

Here in the equation:

Mardia (1970) has shown that for large samples the statistics,

follows a χ2 distribution with p(p+1)(p+2)/6 degrees of freedom and

follows a standard normal distribution. Thus, these two measures allow one to test two hypotheses that are compatible with the assumption of normality.

**Description of Data**

**Dataset 1:**

The first dataset is a retrospective data of 132 type 2 diabetic patients who were diagnosed of diabetes as per American Diabetes Association (ADA) standards, from the data base of Dr. Lal’s Path lab (a reputed NABL certified path lab).

These patients were contacted through a house-to-house survey and up-to-date pathological reports were collected from the time of diagnosis of diabetes till the termination of study.

The data regarding the duration of diabetes and other factors like age at which diabetes was diagnosed;

Fasting Blood Glucose (FBG), Diastolic Blood Pressure (DBP), Systolic Blood Pressure (SBP), Low Density Lipoprotein (LDL) and values of SrCr were recorded for each patient.

In our study, patients with same duration of diabetes have different renal health status. The renal health status of a patient is determined on the basis of SrCr, as the rate of rise in the value of SrCr is an important marker for prediction of

DN.

Thus, using the values of SrCr the data has been classified into two categories namely DN (SrCr ≥ 1.4mg/dl) and non-diabetic nephropathy (NDN) (SrCr < 1.4mg/dl) groups.

**Dataset 2:**

The second dataset consists of 200 type 2 diabetic patients who were again diagnosed as type 2 diabetic as per ADA standards. The pathological reports of these patents were collected from the database from January 2012 to August 2013. As per the availability of information from the collected pathological report of 200 patients, the data regarding the factors FBG, Glycated haemoglobin (HbA1c), SrCr, Age, LDL were recorded foreach patient. From the available reports of 200 patients minimum two reports of FBG and SrCr were available for each patient. But HbA1c and LDL were available for only 92 and 80 patients respectively. Out of 200 patients, only 14 patients were found to exceed the normal range of SrCr (i.e. greater than 1.4 mg/dl) and these were classified as DN patients according to ADA criteria. **APPLICATION: -**

Here 3 tables are created based on thew dataset. Table 1 contains descriptive statistics of 132 type 2 diabetic patients. And Table 2 contains descriptive statistics of 200 type 2 diabetic patients.

For Table 3; AIC (Akaike information criterion is an estimator of in-sample prediction error and thereby relative quality of statistical models for a given set of data. AIC = 2K – 2(log-likelihood).). Here AIC is used to check model’s likelihood against its complexity is used to compare the viability of different parametric models. AIC values with the maximum likelihood estimates (MLE) are presented in table 3.

Table 1

Table 2

Table 3

**Checking of Normal Approximation of Selected Distributions**

The selected Gamma distribution of duration of diabetes (t) tends to normal distribution as its shape parameter is larger than its scale parameter. Hence applying the results from section 2.3 it can be concluded that t follows Normal(17.0400,19.1035) . The normal approximation for t is also judged by graph presented in figure 2. Figure support the claim that normal distribution is good approximation for duration of diabetes. The distribution of SrCr also tends to normal as in case of lognormal distribution the ratio of mean and standard deviation is large (> 6). Applying the results from section 2.3, we can say that SrCr Normal follows (2.3640,0.1796) . The same is depicted by graph presented in figure 3. Hence, all three random variables are marginally normally distributed as FBG was also found to be normally distributed with parameters µ= 170.9643 and σ= 21.4293.

**Figure 2.** Normal approximation for Gamma distribution

**Figure 3.** Normal approximation for Lognormal distribution

**Checking Multivariate Normality of t, SrCr and FBG**

Next, we have checked multivariate normality of t, SrCr and FBG by applying Mardia test. The calculated chi-square (10 degrees of freedom) and standard normal test statistic came out to be 1 κ = 12.4437 and 2 κ = 0.6138 (p-values >0.05) respectively, indicating that t, SrCr and FBG are jointly normally distributed. Now, since the variables: duration of diabetes, SrCr and FBG are marginally normally distributed with significant correlation coefficients (ρt, SrCr = 0.544, ρt, FBG = 0.457, ρSrCr, FBG = 0.697; p-values < 0.0500)and are also jointly normally distributed. Therefore, MVN distribution is a suitable model for representing joint distribution of t, SrCr and FBG and is defined as

##### Estimation of Mean Duration of Diabetes for 60 DN Patients of Dataset 1 by Applying MVN Distribution

##### The primary objective of our study is to estimate the duration of disease for these patients by applying MVN distribution. Since our table 1 consists of 132 type 2 diabetic patients with 60 DN patients and the knowledge of duration of disease and we have first generated a random sample of size 5000 from MVN distribution and estimate the duration of disease for time intervals defined as t ≤ 8, 8 < t ≤ 9, ..., t > 26. To find the mean duration for each time interval we use: -

##### 

##### Here, = mean value of t corresponding to a specific time interval

##### ; mean value of SrCr and FBG

##### ;

##### The procedure of calculation for 60 DN patient of dataset 1:

1. The mean duration of diabetes is calculated from the data of 60 DN patients for only those patients whose mean duration of diabetes is less than or equal to 8 years and is found to be 7.400 years.

##### 2. The SrCr value for these patients range from 1.6500 to 2.2000 mg/dl with mean value 1.925 mg/dl (x SrCr) and FBG value for these patients range from 133 to 199 mg/dl with mean value 166 mg/dl (x FBG).

3. Mean and standard deviation of t (μt, σt), SrCr (μSrCr, σSrCr), FBG ( μFBG, σFBG) and covariance between t, SrCr and FBG (Σ12, Σ12) are calculated from the simulated sample corresponding to the ranges t ≤ 8, 1.6500 ≤ SrCr≤ 2.2000 and 133≤ FBG ≤ 199.

##### 4. Conditional expectation of t|SrCr, FBG is obtained by substituting the above values in equation used in Estimation of Mean Duration of Diabetes for 60 DN Patients of Dataset 1 by Applying MVN Distribution. This gives the mean duration of diabetes for DN patients whose observed duration of disease is less than or equal to 8 years with their mean SrCr and FBG values known. Following the above procedure, the mean durations for all time intervals are estimated and presented in table.

Table 4

##### Estimation of Mean Duration of Diabetes for 14 DN Patients of Dataset 2 by Applying MVN Distribution

Theduration of disease for 14 DN patients of dataset 2 is estimated on the basis of known values of SrCr and FBG.

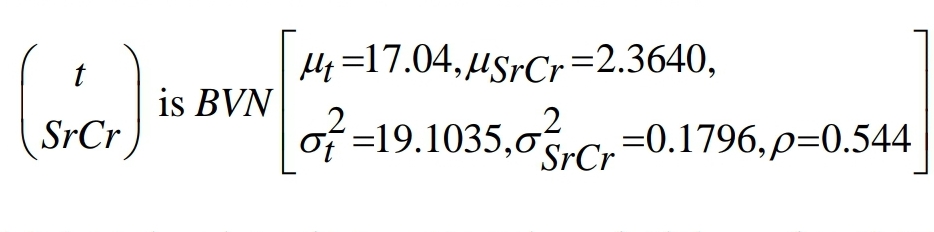
The procedure of calculation for the first DN patient with given mean values of SrCr and FBG is as follows:

1. The mean SrCr and FBG for the first patient is 1.42 mg/dl (SrCr x) and 138 mg/dl (FBG x) respectively.

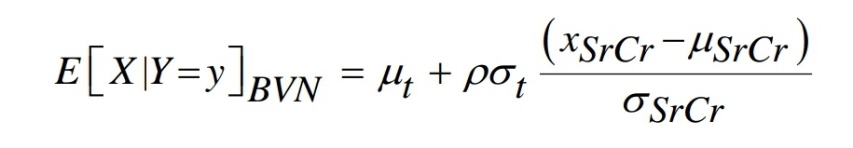
2. Mean and standard deviation of t (μt, σt), SrCr (μSrCr, σSrCr), FBG ( μFBG, σFBG) and covariance between t, SrCr and FBG (Σ12, Σ12) are calculated from the simulated sample corresponding to the ranges 1.4000 ≤ SrCr≤ 1.4200 and 126≤ FBG≤ 138.

3. Conditional expectation of t | SrCr, FBG, is obtained by substituting the above values in equation used in Estimation of Mean Duration of Diabetes for 60 DN Patients of Dataset 1 by Applying MVN Distribution. This gives the mean duration of disease for the first DN patient with known SrCr and FBG values. The mean durations of disease for 14 DN patients are calculated by applying the above procedure and are presented in table 6.

**Bivariate Normal Distribution for Duration of Diabetes (t) and Serum Creatinine (SrCr): -**

We have considered only two random variables viz. duration of diabetes and SrCr. As it was already known that the patients are diabetic and we have estimated the complication arising out of it (using SrCr only). The marginal distributions of these random variables are approximately normal and also, they are jointly normally distributed (applied Mardia test: K1 = (X4)2 = 6.3987 & K2 = 0.7268; p-values >0.05). Thus, 142 BVN distribution is a suitable model for representing joint distribution of t and SrCr and is defined as,

## **Estimation of Mean Duration of Diabetes for 60 DN Patients of Dataset 1 by Applying BVN Distribution**

Firstly, we have estimated the duration of disease for 60 DN patients of dataset 1 by applying conditional expectation under BVN distribution. For this as done for MVN distribution, the data of duration of disease of 60 DN patients is divided into time intervals t≤8,8<t≤9,....,t>26 And then a random sample of size 5000 is generated from BVN distribution with parameters as defined in above equation, and mean duration of disease for each of these intervals are calculated by applying the following equation:

Where μt, σt represents the mean and standard deviation values of t calculated from the simulated sample corresponding to a specific time interval, SrCr x is observed value of SrCr from the data corresponding to a specific interval μSrCr, σSrCr are mean and standard deviation values of SrCr respectively, calculated from the generated sample corresponding to a specific interval. The procedure of calculation for the first-time interval t ≤ 8 is illustrated below:

1. The SrCr value for these patients ranges from 1.6500 to 2.2000 mg/dl with mean value 1.925 mg/dl (xSrCr).
2. Mean and standard deviation of t (μt, σt) and SrCr (μSrCr, σSrCr) are calculated from the simulated sample corresponding to the ranges 1.4000 ≤ SrCr≤ 1.4200.
3. Conditional expectation of t SrCr | is obtained by substituting the above values in equation (9). This gives the mean duration of disease for the first DN patient with known SrCr value.

TABLE: - Estimated mean duration of diabetes of 132 type 2 diabetic nephropathy patients for different time intervals using a generated sample of size 5000 from Bivariate Normal distribution

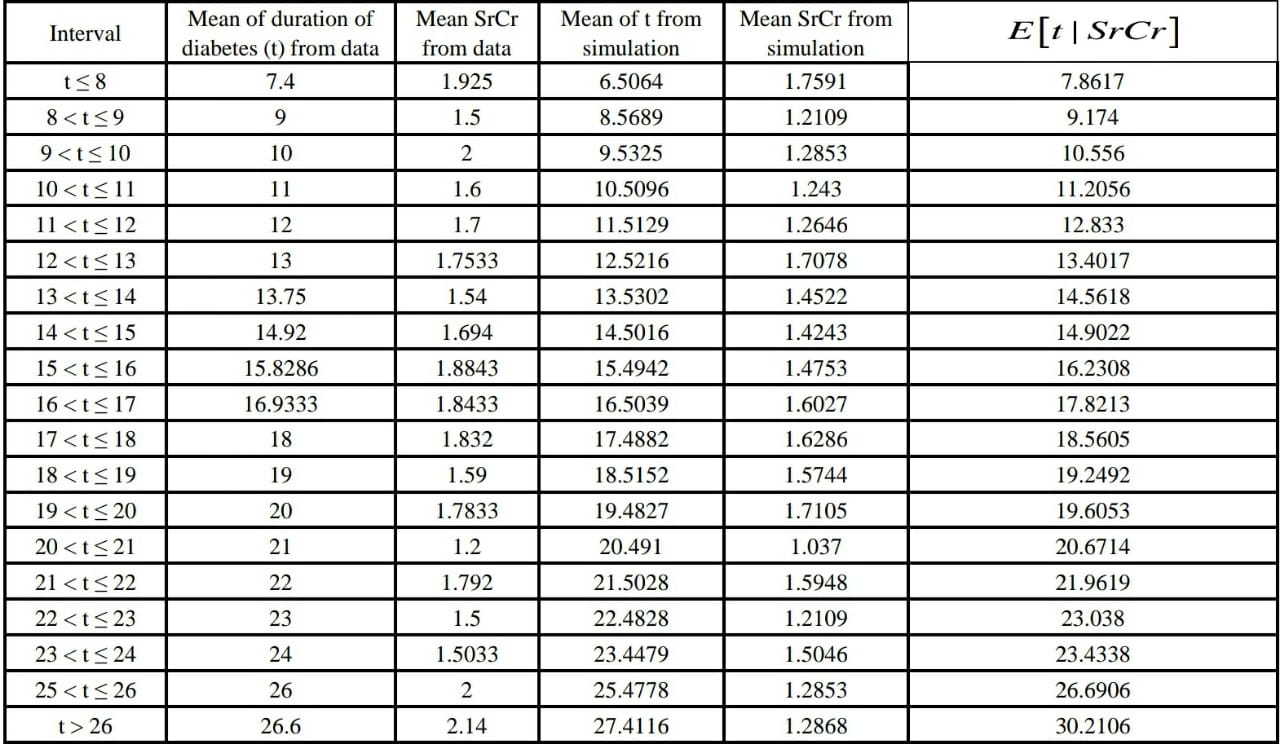
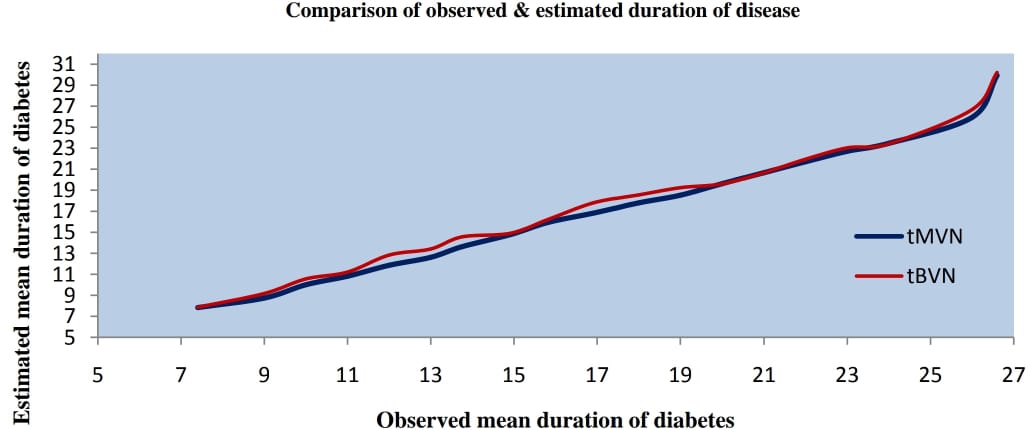
1. The estimated durations of 60 DN patients obtained from MVN and BVN are compared graphically with the observed durations and are presented in below figure

Table 5



**Estimation of Mean Duration of Diabetes for 14 DN of Dataset 2 Patients by Applying BVN Distribution:**

The duration of disease for 14 DN patients of dataset 2 is estimated on the basis of their known value of SrCr. Applying the similar steps as done for MVN case, the

procedure of calculation for the first DN patient with given mean value of SrCr is as follows:

* The mean SrCr for the first patient is 1.42 mg/dl (SrCr x).
* Mean and standard deviation of t (μt, σt) and SrCr (μSrCr, σSrCr) are calculated from the simulated sample corresponding to the ranges 1.4000 ≤ SrCr≤ 1.4200

##### Conditional expectation of t | SrCr is obtained by substituting the above values in equation used for Estimation of Mean Duration of Diabetes for 60 DN Patients of Dataset 1 by Applying BVN Distribution

* This gives the mean duration of disease for the first DN patient with known SrCr value.

The mean duration of disease for 14 DN patients is presented in table 6.

Table 6

These estimated durations of diabetes are further compared graphically with those estimated by applying MVN distribution and are presented in figure below.

**CONCLUSION: -**

From this study we understand the useful approach on the basis of MVN and BVN distributions for estimating the duration of disease. We also concluded that the MVN distribution may be preferred over BVN distribution as more the information the better would be the estimates. The results of the study indicate that the proposed model can contribute meaningfully in dealing with other diabetic complications as well. This study also highlights the usage of simulation when datasets are small.

**OTHER APPLICATION: -**

The MVN Distribution can be used in the algorithms for anomality detection. Animalities in any thing for example animalities in flight engine failures or other things can be estimated.