

Abstract

Through this simulation, we generate a 3 variable - 100 sample multivariate normal data from normal samples. We then check for univariate, bivariate and multivariate normality using several plots.

Libraries

This task has been done in *Python 3.8* and following are the libraries used:

1. Numpy 1.18.1
2. Pandas 1.1.4
3. Matplotlib 3.3.2
4. Scipy 1.5.4
5. Seaborn 0.11.0

Estimated Data Analysis (EDA)

Number of samples (n_samples) = 100

Number of variables (N) = 3

The variables are X_0 , X_1 , X_2

Sample Statistics & Plots

```
[29] samples[:4]
      array([[ -0.44318399,  0.96003095,  1.07110702],
             [ -0.76816421,  0.34016646,  1.35866604],
             [  0.68146524,  0.45377502, -0.96747348],
             [ -1.0636446 ,  1.79069006, -1.06896583]])
```

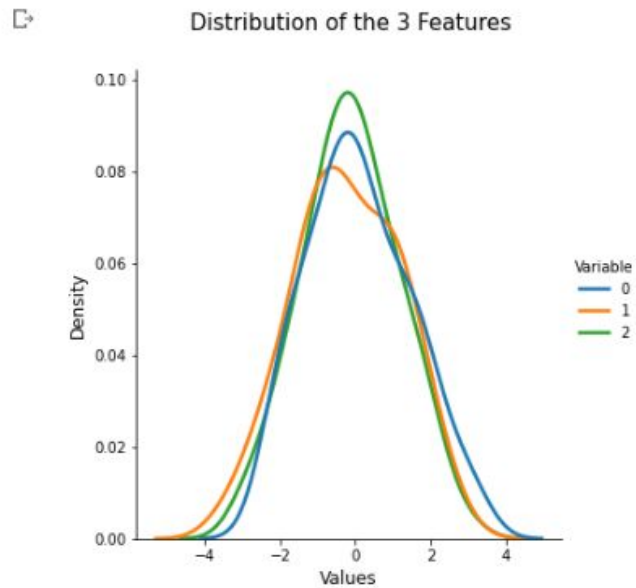
```
samples.shape
```

```
(100, 3)
```

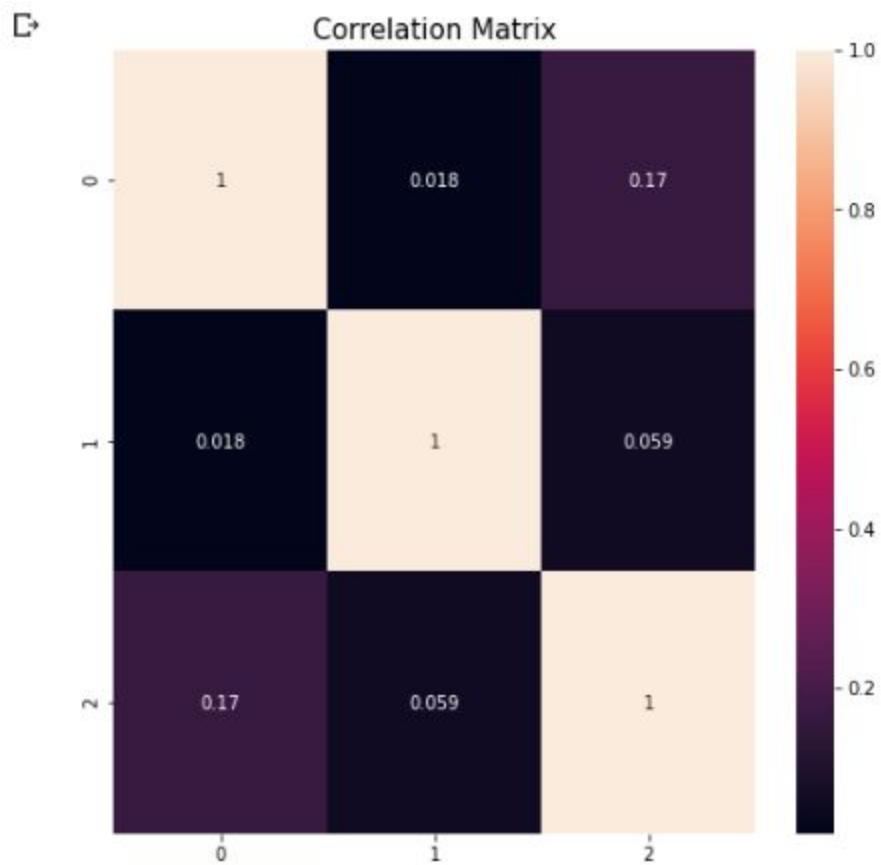
```
[31] np.max(samples, axis=0), np.min(samples, axis=0)
      (array([3.30863127, 2.90324239, 3.05276937]),
       array([-2.45267253, -3.61887696, -3.12483975]))
```

```
[32] np.mean(samples, axis=0), np.std(samples, axis=0)
      (array([ 0.09530217, -0.21527436, -0.12675634]),
       array([1.378316 , 1.41195712, 1.28290428]))
```

From the statistics above, we can see that the (mean, standard deviation) of all the features is approximately (0,1). Hence, we can say that these variables follow standard normal distribution. This can also be proved by the bell-shaped curve of the distributions followed by them.

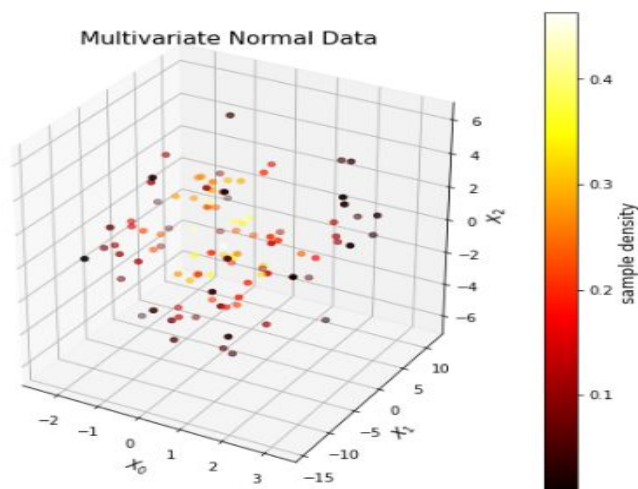


We can also see that the variables have very less correlation and are close to independent variables.



3D scatter plot of sample data

Below is the 3D scatter plot for 100 samples with x-axis, y-axis, z-axis = X_0 , X_1 , X_2 . The color of the points represent the Multivariate Normal Density.



Approach

Sampling Data

For sampling the data, we first generate one random normal sample of size (3,1) : X . We set the mean vector: *mean* to 0 and covariance vector: *covar* to the squared euclidean distances between the points in X .

We then calculate the Cholesky Decomposition matrix: L and generate Independent samples: Z of size (100, 3). The multivariate data is then generated by then scaling the dot product of L and Z with sample mean to fit normal data.

$$Z \cdot \text{dot}(L) + \text{mean}$$

The probability density: *distribution* for these sample points is calculated using the Multivariate Normal Distribution Function.

$$f(x) = (2\pi)^{-0.5p} |\Sigma|^{-0.5} e^{-0.5(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

where,

p = number of variables,
 x = sample vector,
 Σ = covariance matrix,
 μ = mean vection.

Univariate Normality


The normality of univariate data can be proved by plotting the Q-Q plot for each variable. For this we generate the Normal Theoretical Quantiles: *norm_quantiles* using the Normal Percent Point Function. *norm_quantiles* has the size (100,1). Then for each variable, we generate a Scatter Plot with these theoretical values on the x-axis and the sorted variable values on the y-axis.

Theoretically, if the univariate data is normally distributed, the slope of this graph will be close to a straight line. Practically, there will be some outliers. Univariate normality does not mean that the joint distributions of the variables are normally distributed as well. Hence, we now check for bivariate normality.

Bivariate Normality

Bivariate normality can be checked by

1. Scatter plots



2. Plotting contours,
for every pair of the variables. In this case the pairs are (X_0, X_1) , (X_0, X_2) , (X_1, X_2)

If the data indeed is normally distributed, then the contours must be ellipses. Thus, the scatter plots must also confirm this structure. For this, we create a pairwise scatter plot for the variables. This will also show the outliers in the dataset. We can also show that the samples with the highest probability density are concentrated in the center of the contour.

Another way is to plot a graph Chi-Square Q-Q plot for all the bivariate samples. We generate the Mahalanobis Distances for each bivariate sample and plot it against the Theoretical Chi-Square quantiles (generated using Chi Square Percent Point Function). Similar to the univariate Q-Q plot, the data is said to be linear if the slope is close to a straight line.

Multivariate Normality

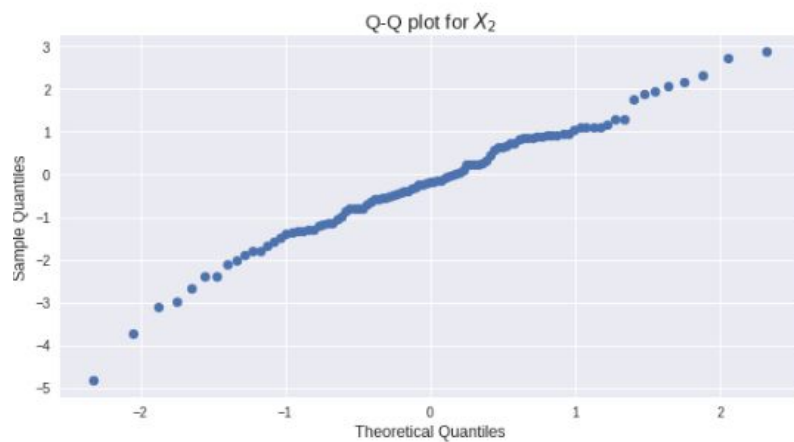
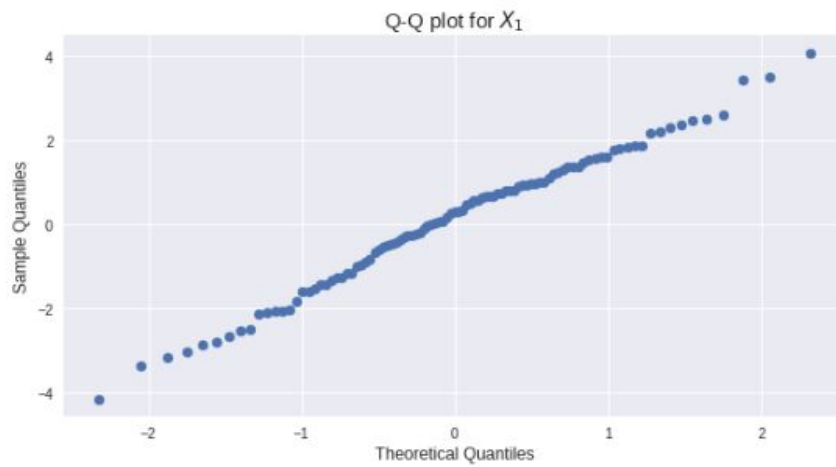
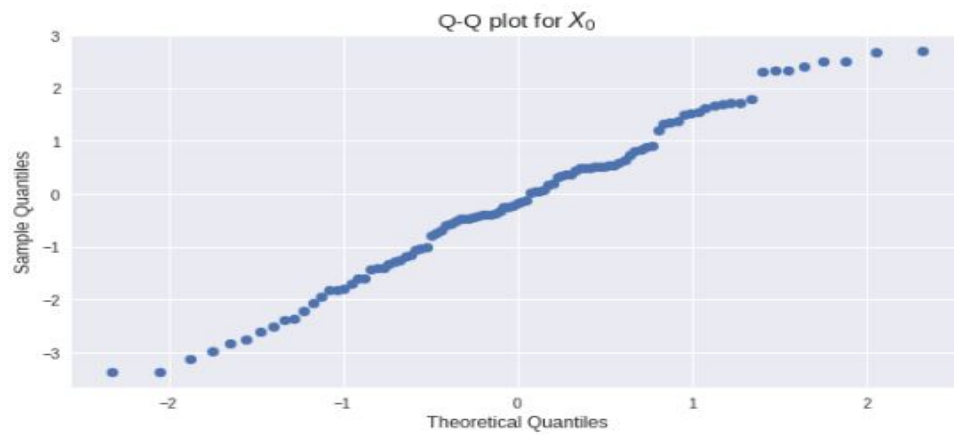
In order to check multivariate normality, we generate a Chi-Square Q-Q plot. For this, we generate the Theoretical Quantiles: *chi_quantiles* of the Chi-Square distribution χ_p^2 . We also generate the squared Mahalanobis distance: *mahalanobis_distances* for each sample. We then scatter plot *chi_quantiles* (x-axis) VS *mahalanobis_distances* (y-axis). If the dataset is indeed normally distributed, then the slope of this graph will be close to a straight line. It will also show the outliers in the dataset, that vary from the general linear trend.

We can also check for normality of the linear combinations of the samples. This will also prove that linear combinations of normal data also follow normal distribution. For this, we plot the same Chi-Square Q-Q plot for the modified sample data.

Results

Univariate Normality

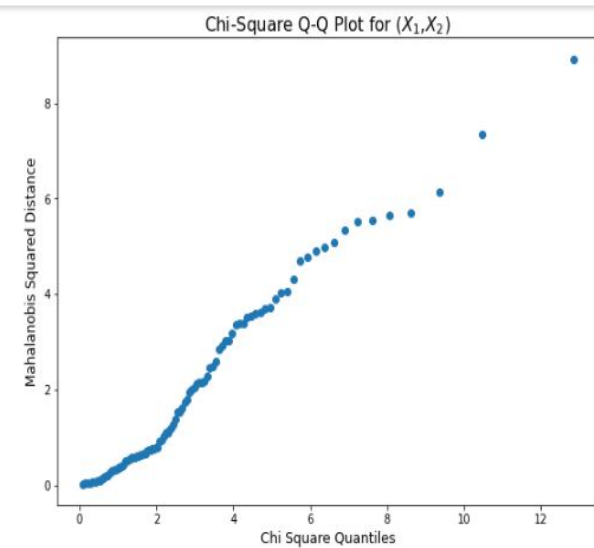
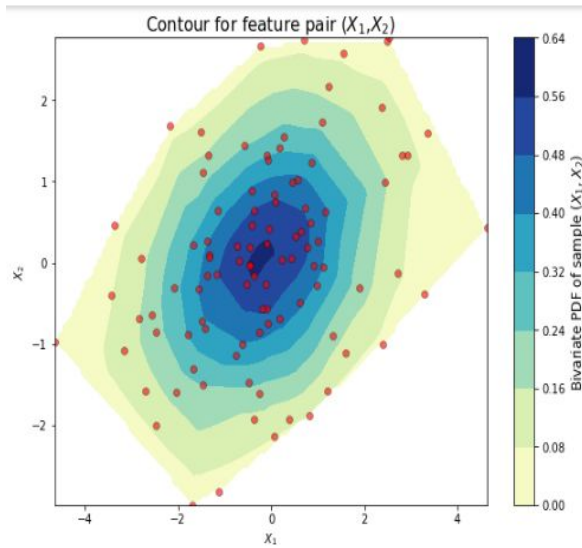
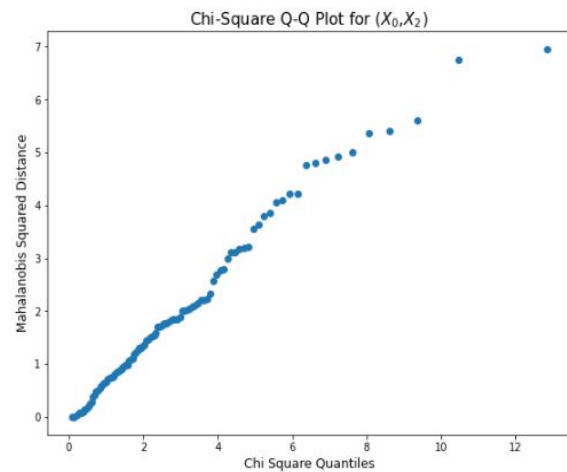
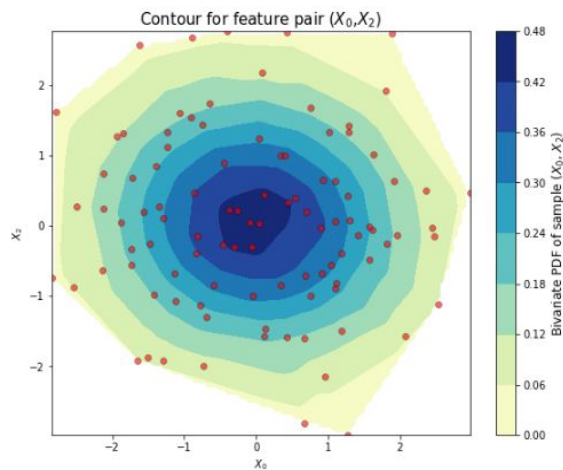
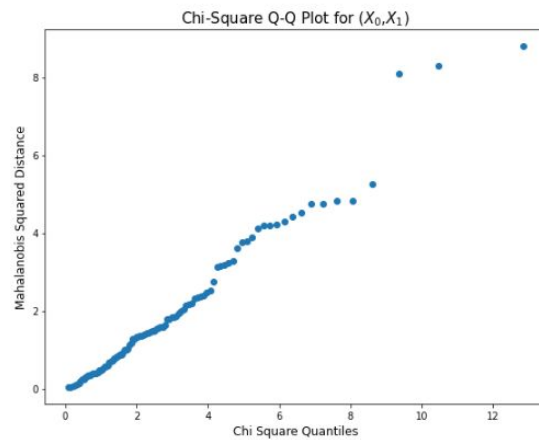
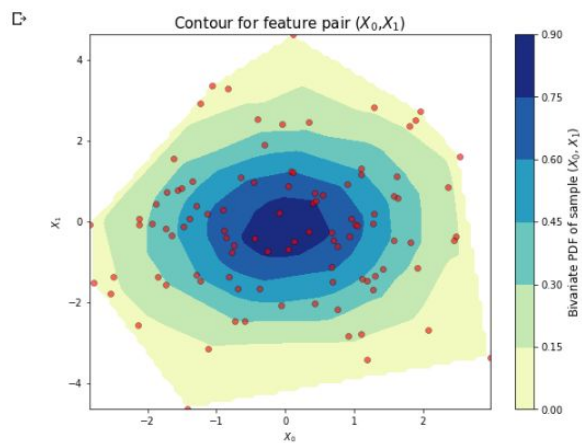
Following are the Q-Q plots for all features- X_0 , X_1 , X_2



The almost straight lines in the graphs show that the features are indeed Univariate.

Bivariate Normality

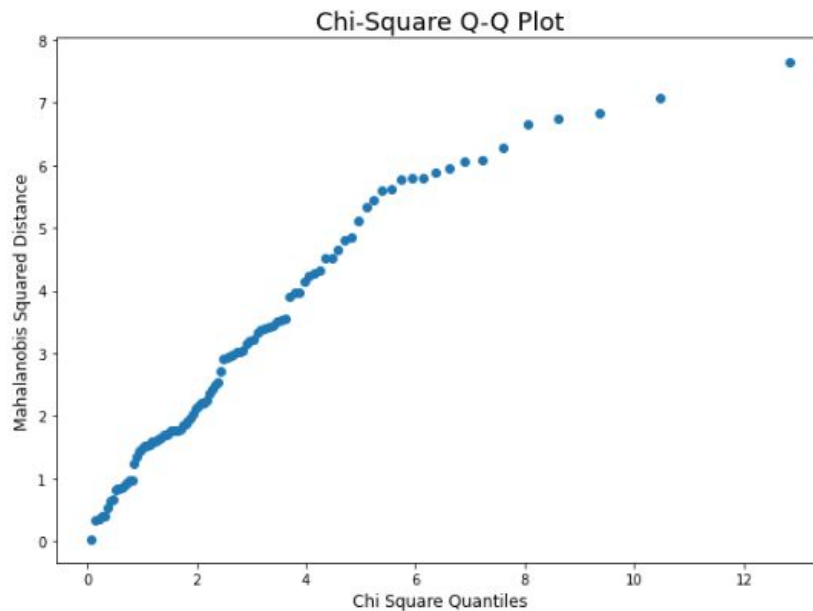
Following are the Contours and Chi-Square Q-Q Plots for the pairs:
 (X_0, X_1) , (X_0, X_2) , (X_1, X_2)



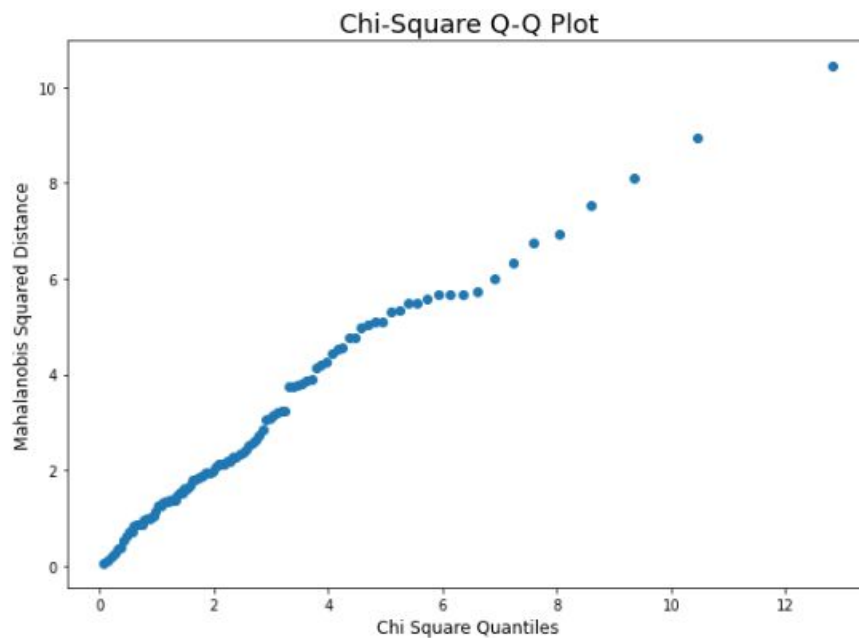
The elliptical structures of the contours along with the almost straight lines in the Q-Q plot prove that the samples are Bivariate.

Multivariate Normality

Following is the Chi-Square Q-Q Plots for the samples.



The almost straight lines in the Q-Q plot prove that the samples are indeed Multivariate. The following Chi-Square Q-Q plot of the modified samples also shows that the linear combinations of the multivariate data is also Multivariate.





Conclusion

From the above methods, we can say that the generated samples are univariate, bivariate and multivariate altogether.