# Chi² Test to Determine the Cut-Off Value for Anomalies Detection with Mahalanobis Distance

**João Felipe de Araújo Caldas[1*], Caique Augusto Cardoso de Moraes[1], Flávio Santos Conterato[1]**
*[1]SENAI CIMATEC University Center; Salvador, Bahia, Brazil*

**This work aims to contemplate the detection of possible anomalies in a dynamic, robust, and effective way using Mahalanobis distance readily with the Chi² test. Tests were performed with p-values of a data set and the cut-off value, both generated by the Chi² test, reliably and dynamically detected possible anomalies. Therefore, this anomaly detection method is more effective for regular anomaly determination based on points farthest from the center.**
**Keywords: Chi². Mahalanobis. Anomalies.**

## Introduction

The use of distance calculation dates back to 650 BC as an attempt to define the shortest possible distance from two points. Thales of Miletus began calculating distances between different objects using two-point references [1,2], for instance, the distance between a ship and shore or height from the top of a pyramid to the example etcetera. Subsequently, Euclidean geometry contributed to calculating the Euclidean distance, and nowadays, we can see its use in the objective life of any person. However, in machine learning, the need to estimate the metric of two points comes from the demand of the data clustering process [1].

Clustering, which is a powerful data mining technique, in turn, is formed by the proximity relationship of the defined methods, being able to be distinguished between two types by the distance of the defined points, using the defined formulas and mathematics methods, or by the degree of similarity based on their characteristics [3].

There are some measures to determine the interval between points; among them are Hamming distance, the Minkowski distance, the Manhattan distance, and the most famous, the Euclidean distance [4]. Even so, the article's primary subject was the statistical measure created by the Indian scientist Prasanta Chandra Mahalanobis, called Mahalanobis distance, used in the experiment [5.

The $\chi^2$(chi-square), or simply Chi²(chi-square) test, is described as a non-parametric test; that is, it does not depend on population parameters; in addition to being a hypothesis test, the principle basic to this test is to compare proportions, that is, possible divergences between the observed and expected frequencies for a specific event [6].

An example of Chi² application was the usability to assess the significance of each parameter considered for risk stratification and mortality prediction for definitive values about COVID-19 [7], dealing directly with data processing to improve the final results. Mahalanobis distance is an effective multivariate distance metric that measures the distance between a point (vector) and a distribution [5,6] as per Formula 1. This metric is a multivariate distance calculation metric, efficiently measuring spatial distributions between points and vector distributions at intervals, thus having prominence in anomaly detection applications, classification of untreated datasets, and in some cases in classifications of classes and most ignored use cases [5].

The formula for the Mahalanobis distance is given by [6]:

$$D^2 = (x - t)^T \cdot C^{-1} \cdot (x - m) \qquad (1)$$

In which, D² is the square of the Mahalanobis distance; x is the observation vector; m is the

Received on 12 December 2022; revised 10 February 2023.
Address for correspondence: João Felipe de Araújo Caldas. Conjunto Jardim das Limeiras, 32 - São Marcos, Salvador, Bahia, Brazil | Zipcode: 41250-440. E-mail: jfdac11@gmail.com.

J Bioeng. Tech. Health 2023;6(1):58-61

vector of mean values of the independent variables; $C^{-1}$ is the inverse of the covariance matrix of the independent variables.

Formula 1 also has some observations; if the matrix (x - m) is diagonal, then the distance measure is analogous to the normalized Euclidean distance [5].

This article proposes to showcase the efficacy of anomaly detection in conjunction with the distance between a distribution and Mahalanobis points. Arguments supporting the use of this technique will be presented in dissonance with the current literature approach [8,9] regarding anomaly detection in a database. After experimentation and comparisons with the Chi² test, the x values farthest from a centroid-based method lacked dynamism and showed high chances of detecting false positives. Hence, the correlation between Chi² and Mahalanobis distance is displayed to evidence the accuracy of the cut-off values for anomaly detection proposed by the result of the Chi² test (or p-value) present in this paper.

## Materials and Methods

The dataset used was obtained from the learning and challenge platform Kaggle [10], a well-known platform in the Data Science area; the dataset is called World Happiness Report, which contains information about world happiness. There are two main reasons for the choice of the dataset. The first is that the dataset needs to have a high degree of relationship, i.e., that presents a precise functioning with highly correlated data in contrast to neural networks and especially in Pearson functional networks [11] so that the Chi² can stand out in the results of possible anomalies to facilitate the understanding and comprehension of the content, the second reason was that after the pandemic, socioeconomic behaviors were impacted, such as happiness and quality of life [12], for these factors the database proved to be more attractive, thus hoping to return not only anomalous results but also the difference in the degree of happiness over time. The dataset has the period from 2015 to 2022; however, it was used from 2020 to 2022 to detect possible anomalies, considering that the COVID-19 virus started spreading in January 2020 and causing the pandemic in March 2020 [13] and that this pandemic impacts to society [12]. That had as motivation for the search for the possible impacts of the COVID-19 pandemic on world happiness. During the development, it was verified that the columns with higher correlation using the Pearson correlation coefficient resulted in the upper and lower whisker columns with 99% correlation between them, thus, being chosen by the high degree of correlation to the experiment. Subsequently, the clustering was performed by defining only one cluster (group), the calculation of the data centroid, the computation of the Mahalanobis distance for each point, and the Chi² test were performed with the help of the scipy library [14], and using the cdf function (cumulative distribution function).

Consequently, finding the p-value of each point was possible to establish the cut-off values for anomaly determination. Then the possible anomalies were separated. All tests and experiments were done using Python version 3.7.10 [15].

Chi² is used as a statistical hypothesis test and as a cut-off value, i.e., a value to determine whether a piece of data is anomalous or not. The statistical hypothesis test is performed because the Mahalanobis distance results in the squared distance, the 'D²' [16].

To find the cut-off value, one needs to know the limits of the grouping of the data and the degree of freedom (c) will be the number of variables/ columns present in the database [11]. This calculation is represented by Formula 2, which is given by the formula in which the variable O is the observed value and E is the expected value.

$$X_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

Mahalanobis distance is closely related to Chi² in the way that [6,18] proposed: the squared Mahalanobis distance of a Gaussian (normal) distribution is distributed by Chi². Therefore, the

Chi² test will result in a variable called p-values, intended to demonstrate whether the test results are quantitatively significant. The following considerations are essential to perform a Chi² test and obtain the p-values [6]:

- The degree of freedom is defined as the number of categories minus 1 [6].
- The tester chooses the alpha level (α). Usually, the alpha level is 0.05 (5%), but one can also have other levels, such as 0.01 or 0.10 [6].

The advantages of Chi² include its robustness in data distribution, easy calculation, and detailed information that can be obtained from testing. The caveat to using this methodology in studies is that they cannot meet the parameter assumptions and flexibility in handling two aspects of group data and multi-group studies [16].

## Results and Discussion

After the experiments had been carried out, the result of the detection of anomalies was compared using p-values and the cut-off value of the Chi² hypothesis test. Figure 1 shows three ellipses based on the Chi² test for the database chosen. The values outside the limit of the first ellipse (blue) are considered anomalies (annotated with their respective index). In this case, the delimitation of the ellipse is the critical cut-off value found by performing the Chi² test. It is worth pointing out that the alpha level value was set to 0.01. The standard deviation of the first ellipse measures the ellipses in red and yellow. For example, the red ellipse is the standard deviation multiplied by two, and the yellow is multiplied by three.

In Figure 2, the ellipse continued to be made similarly, but the cut-off value was determined by the p-values also generated by the Chi² test. P-values that have a significance level lower than 0.01 are considered anomalies. Therefore, both methods can be used equally for the same purpose. However, to find the p-value, only one calculation is needed, while the critical cut-off values of the Chi² test are found after more calculations. Depending on the alpha value used in the Chi² test, the result of possible anomalies may be equal to the result of using p-values as a cut-off value. In the case of this database, when using the alpha value as 0.001, the result of both methods are the same. The critical value used in the first method and the p-value of the second method are two different approaches to the same result [17].

**Figure 1.** Chi² critical value cut-off.

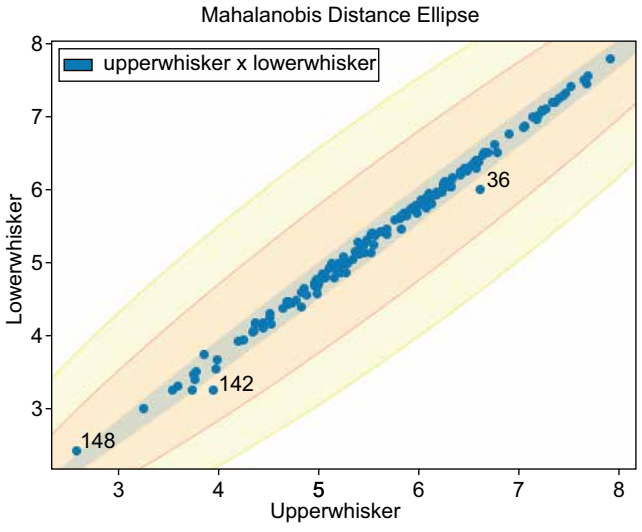
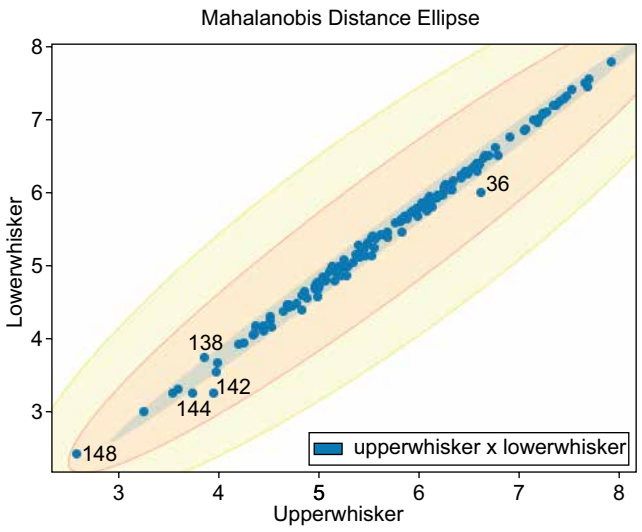



**Figure 2.** Chi² p-value cut-off.



## Conclusion

The anomaly detection is successfully performed, and the need to streamline it is successfully met. The values determined as possible anomalies are established according to a cut-off value calculated by hypothesis testing, unlike how it is commonly performed, measuring a certain amount of points furthest from the centroid.

The results collected in this research are of paramount importance to detecting anomalies since they demonstrate the detection of anomalies through a specific cut-off value, that is, dynamically and more reliably. Therefore, verifying and testing the effectiveness of Chi² with other distance metrics or with many categories is recommended for future work.

## Acknowledgments

We thank Conterato Analytics and Aton Engenharia for supporting and directing the work done and for the opportunity for growth and learning.

## References

1. Rezende EQF, Queiroz MLB. Geometria Euclidiana Plana e Construções Geométricas. Campinas: Editora UNICAMP, 2000.
2. Dante LR. Matemática: Contextos & Aplicações - Volume 1. São Paulo: Editora Ática, 2011.
3. Pinto R. Entendendo porque é que a distância certa faz toda a diferença. 2020. Available:https://medium.com/data-hackers/entendendo-porque-%C3%A9-que-a-dist%C3%A2ncia-certa-faz-toda-a-diferen%C3%A7a-648030c9bae2. Accessed: Jun 23, 2022.
4. Brownlee J. Distance Measures for Machine Learning. 2020.Available:https://machinelearningmastery.com/distance-measures-for-machine-learning/#:~:text=Perhaps%20four%20of%20the%20most,Manhattan%20Distance. Accessed: Jun 25, 2022.
5. Prabhakaran S.Mahalanobis Distance – Understanding the math with examples (python). 2019. Available:https://www.machinelearningplus.com/statistics/mahalanobis-distance/. Accessed: Jun 24, 2022.
6. Glen S. Chi-Square Statistic: How to Calculate It / Distribution. www.StatisticsHowTo.com. Available:https://www.statisticshowto.com/probability-and-statistics/chi-square/. Accessed: May 7, 2021.
7. Alle S et al. COVID-19 Risk stratification and mortality prediction in hospitalized Indian patients: Harnessing clinical data for public health benefits. PloS One2022;17(3):e0264785.
8. Data Tech Notes. Anomaly Detection Example with K-means in Python. May 13, 2020. Available from:https://www.datatechnotes.com/2020/05/anomaly-detection-with-kmeans-in-python.html. Accessed: Sept 10, 2022.
9. Dino L. Outlier detection using K-means clustering in python. April 19, 2022. Available from: https://towardsdev.com/outlier-detection-using-k-meansclustering-in-python-214188fc90e8. Accessed: Sept 10, 2022.
10. World Happiness Report up to 2022. Kaggle, 2022. Available:https://www.kaggle.com/datasets/mathurinache/world-happiness-report. Accessed: Jul 1, 2022.
11. Ivanov AI, Vyatchanin SE, Lozhnikov PS. Comparable estimation of network power for chi-squared Pearson functional networks and Bayes hyperbolic functional networks while processing biometric data. In: 2017 International Siberian Conference on Control and Communications (SIBCON). IEEE 2017:1-3.
12. Qiu W et al. The pandemic and its impacts. Health, Culture and Society. 2017;9:1-11.
13. Strabelli TMV, Uip DE. COVID-19 e o Coração. Arquivos Brasileiros de Cardiologia 2020;114:598-600.
14. Scipy. scipy.stats.chi2. 2022. Available:https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2.html. Accessed: Jun 30, 2022.
15. Python, Python Release, 2022, Available: https://www.python.org/downloads/release/python-3710/. Accessed: 30 Jun, 2022.
16. McHugh ML. The chi-square test of independence. Biochemia medica 2013;23(2):143-149.
18. Glen S. P-value *vs*. critical value. Jul 26, 2020. Available from:https://www.datasciencecentral.com/p-value-vs-critical-value/. Accessed: Sept 9, 2022.
19. Thill M. The relationship between the Mahalanobis distance and the Chi-squared distribution. ML & Stats. 2017. Available:https://markusthill.github.io/mahalanbis-chi-squared/. Accessed: May 7, 2021.
20. Cansiz S. Multivariate outlier detection in python. towards data. Science. 2020. Available:https://towardsdatascience.com/multivariate-outlier-detection-in-python-e946cfc843b3. Accessed: May 3, 2021.