How can the Mahalanobis Distance be calculated in R?

Authored by

stats writer

April 18, 2024

RECOMMENDED CITATION

stats writer (2024). *How can the Mahalanobis Distance be calculated in R?*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=136501

ARABPSYCHOLOGY.COM

The Mahalanobis Distance is a statistical measure used to calculate the distance between two points in a multivariate data set. In R, this distance can be calculated by using the "mahalanobis" function from the stats package. This function takes in the data points and their corresponding covariance matrix as inputs and outputs a numerical value representing the distance. It uses a formula that takes into account the correlation between variables, making it a more accurate measure than other distance metrics. This allows for a better understanding of the relationship between data points and can be useful in various statistical analyses, such as clustering and outlier detection. By utilizing the "mahalanobis" function in R, researchers and analysts can easily and efficiently calculate the Mahalanobis Distance and incorporate it into their data analysis workflows.

Calculate Mahalanobis Distance in R

The Mahalanobis distance is the distance between two points in a multivariate space.

It is often used to find outliers in statistical analyses that involve several variables.

This tutorial explains how to calculate the Mahalanobis distance in R.

Example: Mahalanobis Distance in R

Use the following steps to calculate the Mahalanobis distance for every in a dataset in R.

Step 1: Create the dataset.

First, we'll create a dataset that displays the exam score of 20 students along with the number of hours they

ARABPSYCHOLOGY.COM

spent studying, the number of prep exams they took, and their current grade in the course:

```
#create data
df = data.frame(score = c(91, 93, 72, 87, 86, 73, 68, 87, 78, 99, 95, 76, 84, 96, 76, 80, 83, 84, 73, 74),
hours = c(16, 6, 3, 1, 2, 3, 2, 5, 2, 5, 2, 3, 4, 3, 3, 3, 4, 3, 4, 4),
prep = c(3, 4, 0, 3, 4, 0, 1, 2, 1, 2, 3, 3, 3, 2, 2, 2, 3, 3, 2, 2),
grade = c(70, 88, 80, 83, 88, 84, 78, 94, 90, 93, 89, 82, 95, 94, 81, 93, 93, 90, 89, 89))

#view first six rows of data
head(df)

score hours prep grade
1 91 16 3 70
2 93 6 4 88
3 72 3 0 80
4 87 1 3 83
5 86 2 4 88
6 73 3 0 84
```

Step 2: Calculate the Mahalanobis distance for each observation.

ARABPSYCHOLOGY.COM

Next, we'll use the built-in mahalanobis() function in R to calculate the Mahalanobis distance for each observation, which uses the following syntax:

mahalanobis(x, center, cov)

where:

x: matrix of datacenter: mean vector of the distributioncov: covariance matrix of the distribution

The following code shows how to implement this function for our dataset:

#calculate Mahalanobis distance for each observation
mahalanobis(df, colMeans(df), cov(df))

16.5019630 2.6392864 4.8507973 5.2012612 3.8287341 4.0905633
4.2836303 2.4198736 1.6519576 5.6578253 3.9658770 2.9350178
2.8102109 4.3682945 1.5610165 1.4595069 2.0245748 0.7502536
2.7351292 2.2642268

Step 3: Calculate the p-value for each Mahalanobis

ARABPSYCHOLOGY.COM

distance.

We can see that some of the Mahalanobis distances are much larger than others.

To determine if any of the distances are statistically significant, we need to calculate their .

The p-value for each distance is calculated as the p-value that corresponds to the Chi-Square statistic of the Mahalanobis distance with k-1 degrees of freedom, where k = number of variables.

So, in this case we'll use a degrees of freedom of 4-1 = 3.

```
#create new column in data frame to hold Mahalanobis distances
df$mahal <- mahalanobis(df, colMeans(df), cov(df))

#create new column in data frame to hold p-value for each Mahalanobis distance
df$p <- pchisq(df$mahal, df=3, lower.tail=FALSE)

#view data frame
df
```

ARABPSYCHOLOGY.COM

score	hours	prep	grade	mahal	p
1	91	16	3	70	16.5019630	0.0008945642
2	93	6	4	88	2.6392864	0.4506437265
3	72	3	0	80	4.8507973	0.1830542407
4	87	1	3	83	5.2012612	0.1576392526
5	86	2	4	88	3.8287341	0.2805615121
6	73	3	0	84	4.0905633	0.2518495222
7	68	2	1	78	4.2836303	0.2324211504
8	87	5	2	94	2.4198736	0.4899458807
9	78	2	1	90	1.6519576	0.6476670033
10	99	5	2	93	5.6578253	0.1294978092
11	95	2	3	89	3.9658770	0.2651724541
12	76	3	3	82	2.9350178	0.4017530495
13	84	4	3	95	2.8102109	0.4218217836
14	96	3	2	94	4.3682945	0.2243432904
15	76	3	2	81	1.5610165	0.6682610031
16	80	3	2	93	1.4595069	0.6916471506
17	83	4	3	93	2.0245748	0.5673218169
18	84	3	3	90	0.7502536	0.8613248635
19	73	4	2	89	2.7351292	0.4342904353
20	74	4	2	89	2.2642268	0.5194087143

Typically a p-value that is less than .001 is considered to be an outlier.

ARABPSYCHOLOGY.COM

We can see that the first observation is an outlier in the dataset because it has a p-value less than .001.

Depending on the context of the problem, you may decide to remove this observation from the dataset since it's an outlier and could affect the results of the analysis.

How to Perform Multivariate Normality Tests in R

ARABPSYCHOLOGY.COM