How to Calculate Mahalanobis Distance in SPSS

Authored by
stats writer

December 25, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Calculate Mahalanobis Distance in SPSS*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=108805

The Mahalanobis distance is a powerful statistical measure developed by P. C. Mahalanobis in 1936. Unlike Euclidean distance, which measures distance in absolute terms, the Mahalanobis distance accounts for the covariance structure of the data, thereby measuring the separation between a point and a distribution, or between two distributions. This makes it particularly useful for analyzing complex datasets where variables are correlated.

In the context of statistical modeling, the primary application of Mahalanobis distance is the rigorous identification of outliers, particularly multivariate outliers. A multivariate outlier is an observation that is extreme not necessarily on any single variable, but rather in the combination of its variable scores across a multivariate space. Identifying and handling these unusual observations is crucial, as they can significantly bias regression coefficients and distort the results of inferential statistics.

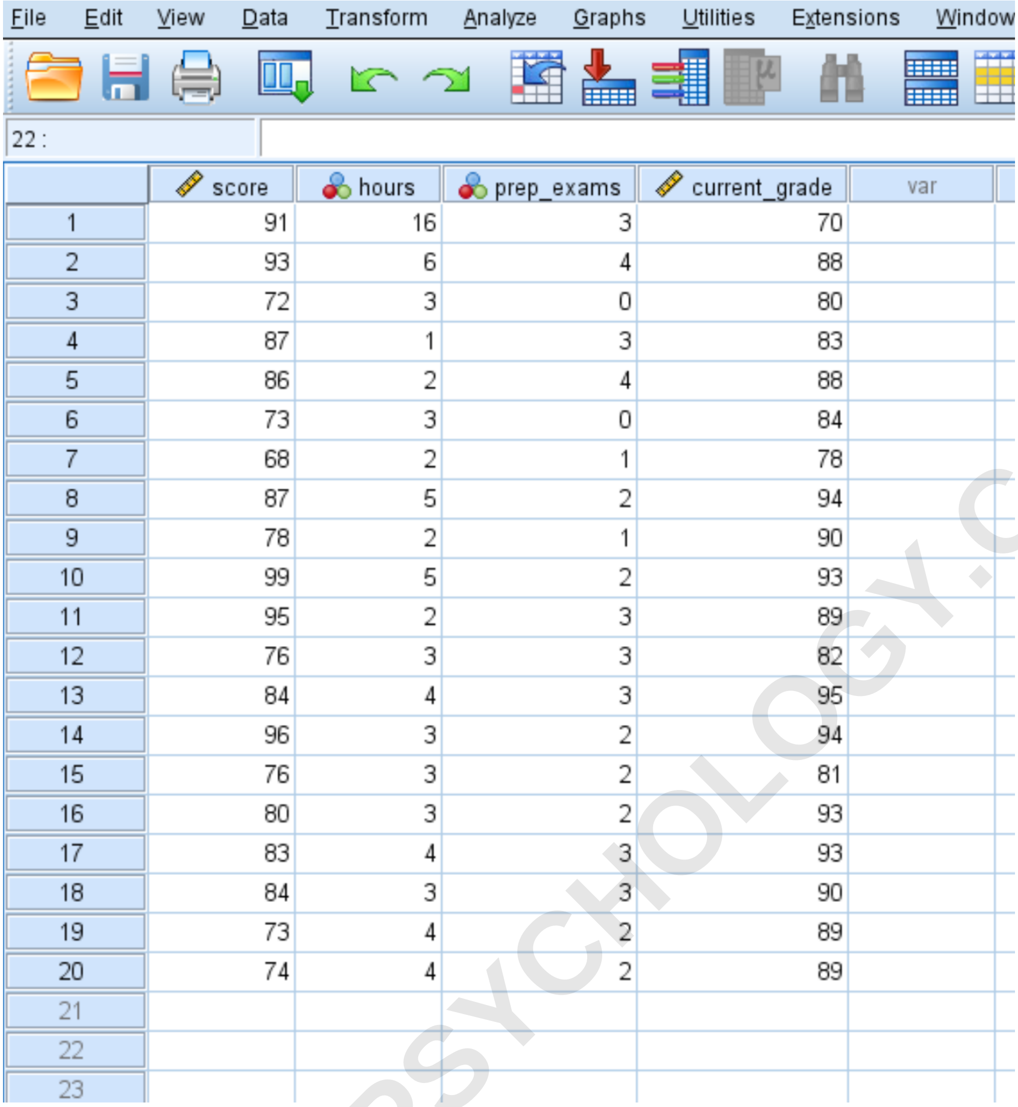
This comprehensive guide details the precise procedure for calculating and interpreting the Mahalanobis distance using SPSS (Statistical Package for the Social Sciences). We will walk through a practical example, generating the distance measures and calculating the corresponding p-values necessary for robust outlier detection.

Illustrative Example: Detecting Multivariate Outliers in a Student Dataset

To demonstrate the utility of the Mahalanobis distance, consider a common scenario in educational research. Suppose we have collected data from 20 students, tracking several key performance indicators (KPIs). Our goal is to assess whether any student exhibits a combination of scores that is statistically unusual compared to the overall group distribution.

The dataset includes four critical variables: the dependent variable, **Exam Score**, and three predictor variables: **Hours Studied**, **Number of Prep Exams Taken**, and **Current Course Grade**. Calculating the Mahalanobis distance for each observation allows us to determine if any student's multivariate profile deviates significantly from the mean profile of the 20 students, thereby identifying potential outliers.

The following image displays the structure of the dataset as entered into SPSS. We will use these data points throughout the subsequent steps to calculate the distance measure for each individual observation.



	score	hours	prep_exams	current_grade	var
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	
7	68	2	1	78	
8	87	5	2	94	
9	78	2	1	90	
10	99	5	2	93	
11	95	2	3	89	
12	76	3	3	82	
13	84	4	3	95	
14	96	3	2	94	
15	76	3	2	81	
16	80	3	2	93	
17	83	4	3	93	
18	84	3	3	90	
19	73	4	2	89	
20	74	4	2	89	
21					
22					
23					

We will now follow a standardized series of steps within the SPSS environment to calculate and validate the Mahalanobis distance for every data point, ultimately identifying any observation that qualifies as a significant multivariate outlier.

Procedure for Calculating Mahalanobis Distance in SPSS

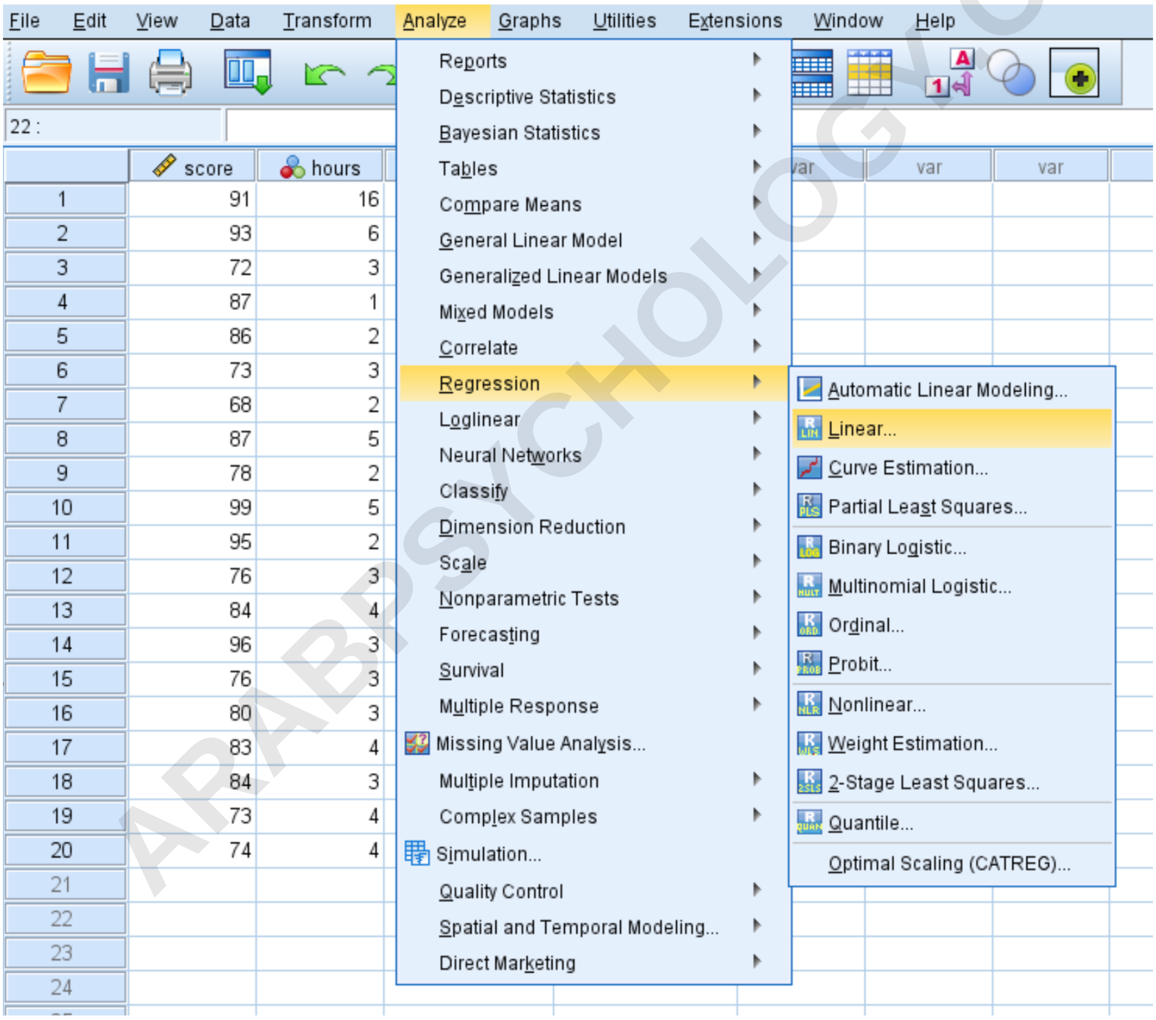
The SPSS software calculates the Mahalanobis distance as part of the regression analysis module. Although we are primarily interested in the distance metric itself, we must initiate a linear regression procedure to access the necessary diagnostic statistics. Follow these precise steps to generate the distance scores for every observation in your dataset.

The methodology leverages the fact that the Mahalanobis measure depends on the pooled variance-covariance matrix of the predictor variables, which is internally computed during the setup of a multivariate analysis like linear regression.

Step 1: Accessing the Linear Regression Dialogue Box

The initial step requires navigating to the appropriate statistical procedure within the SPSS interface. Click on the **Analyze** tab in the main menu. From the drop-down menu, select **Regression**, and then choose **Linear...**. This action opens the primary dialogue box where variable assignments are made for the regression model.

In this dialogue box, you must define the dependent and independent variables. Remember that while the Mahalanobis distance calculation is independent of the regression outcome, the metric relies on the variables included in the model to determine the covariance matrix. Thus, all variables relevant to the outlier detection must be included.



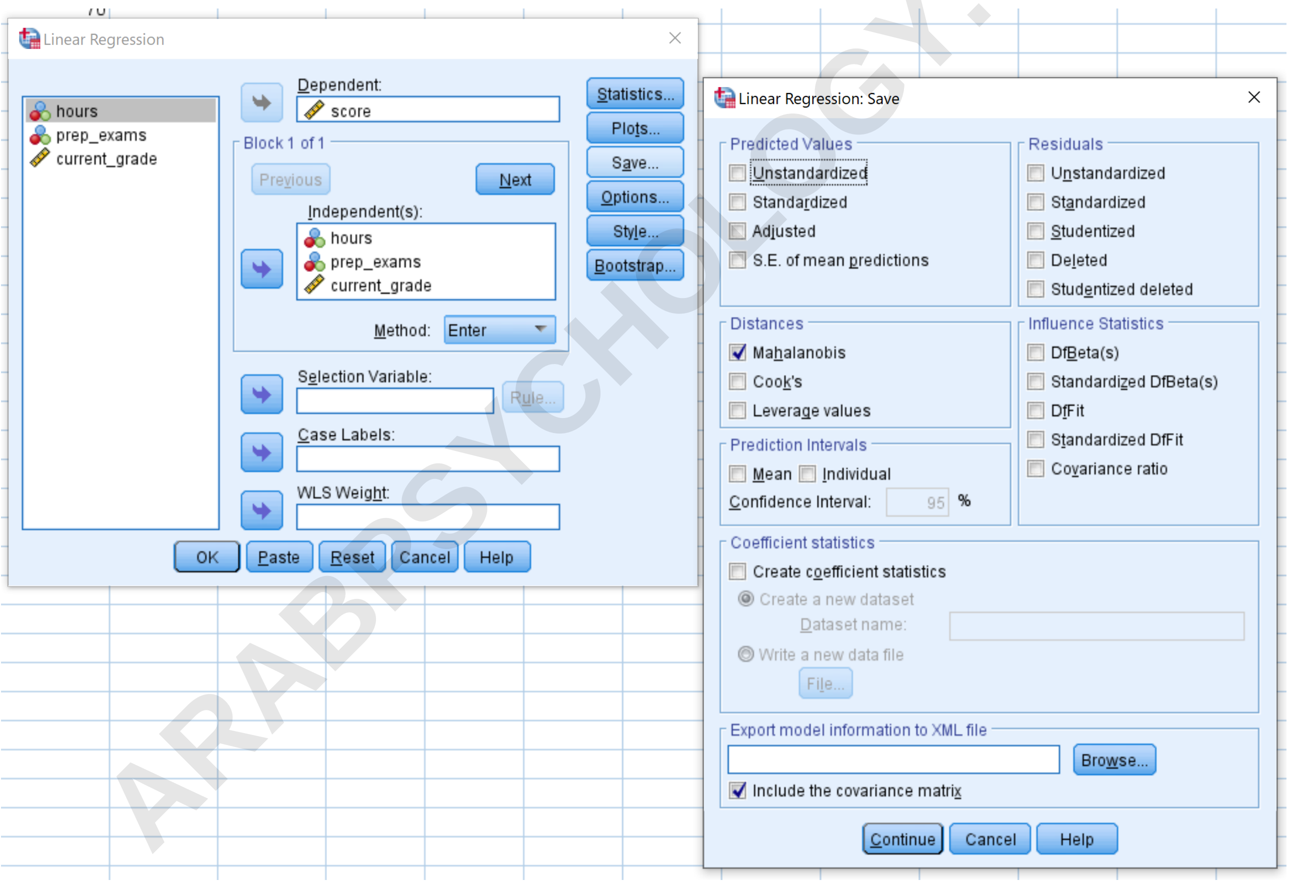
Step 2: Configuring Diagnostic Statistics and Saving the Distance

Once the Linear Regression dialogue box is open, proceed to assign your variables. Drag the outcome variable (**Score**) into the box labeled **Dependent**. Next, drag all three predictor variables

(**Hours Studied**, **Prep Exams Taken**, and **Current Grade**) into the box labeled **Independent(s)**.

Crucially, before running the analysis, click the **Save** button located near the bottom of the Linear Regression window. A new auxiliary window will appear, presenting various options for saving diagnostic statistics and residuals to the dataset. Within the section labeled 'Distances', ensure that the checkbox next to **Mahalanobis** is emphatically checked. This tells SPSS to compute and save the distance metric for each case.

After confirming the selection, click **Continue** to close the auxiliary window, and then click **OK** in the main Linear Regression dialogue box. SPSS will execute the regression analysis and, more importantly for our goal, append the calculated Mahalanobis distance values directly into a new column in your dataset.



Reviewing and Interpreting Raw Mahalanobis Scores

Upon successful execution of Step 2, a new variable named **MAH_1** will be generated and added as the last column in your SPSS Data View. This column contains the raw Mahalanobis distance score for every individual observation. These scores represent how far each data point lies from the centroid (the multivariate mean) of the dataset, taking correlations between variables into account.

A high Mahalanobis score indicates that the observation is far from the center of the multivariate distribution. While simply observing high scores can give a preliminary idea of potential outliers, this raw score alone does not confirm statistical significance. We note that some distances are visibly larger than others, prompting the need for further statistical evaluation.

	score	hours	prep_exams	current_grade	MAH_1
1	91	16	3	70	16.41507
2	93	6	4	88	2.58699
3	72	3	0	80	4.74810
4	87	1	3	83	3.80675
5	86	2	4	88	3.80845
6	73	3	0	84	4.07914
7	68	2	1	78	4.06957
8	87	5	2	94	2.41850
9	78	2	1	90	1.65188
10	99	5	2	93	1.92343
11	95	2	3	89	1.08859
12	76	3	3	82	2.28033
13	84	4	3	95	1.92415
14	96	3	2	94	1.36636
15	76	3	2	81	1.52369
16	80	3	2	93	1.00037
17	83	4	3	93	1.19153
18	84	3	3	90	.66156
19	73	4	2	89	.22777
20	74	4	2	89	.22777

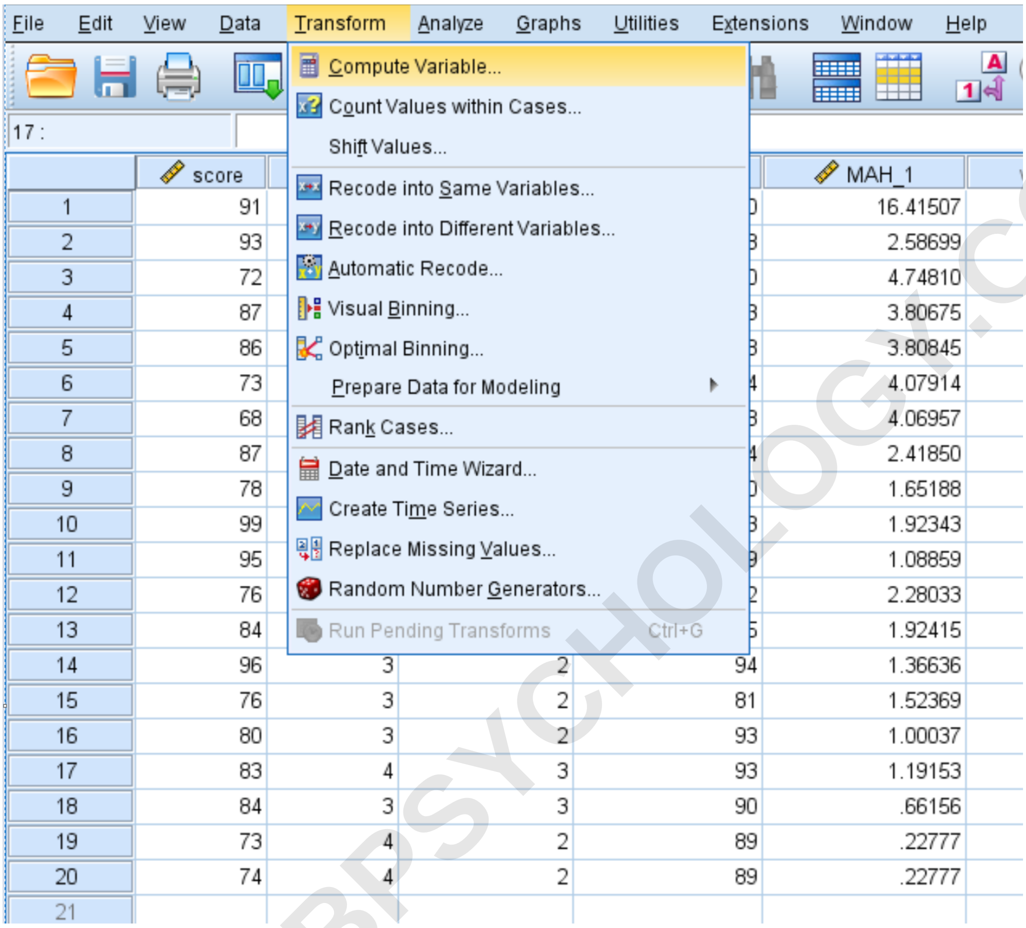
To rigorously determine if any of these distances are statistically significant--meaning the observations are true multivariate outliers--we must calculate the corresponding p-value for each distance measure. The Mahalanobis distance follows a Chi-Square distribution with degrees of freedom equal to the number of predictor variables, which allows us to convert the distance score into a probability value.

Step 3: Calculating the P-Values for Statistical Significance

The process of converting the Mahalanobis distance (MAH_1) into a p-value requires using the data transformation capabilities within SPSS. This procedure utilizes the Cumulative Distribution Function (CDF) of the Chi-Square distribution to determine the probability of observing a distance this extreme or greater.

Initiating the Compute Variable Function

Begin by navigating to the **Transform** tab in the main menu, and then select **Compute Variable...**. This opens the dialogue box necessary for creating new calculated variables based on existing data. This is where we will apply the statistical formula.



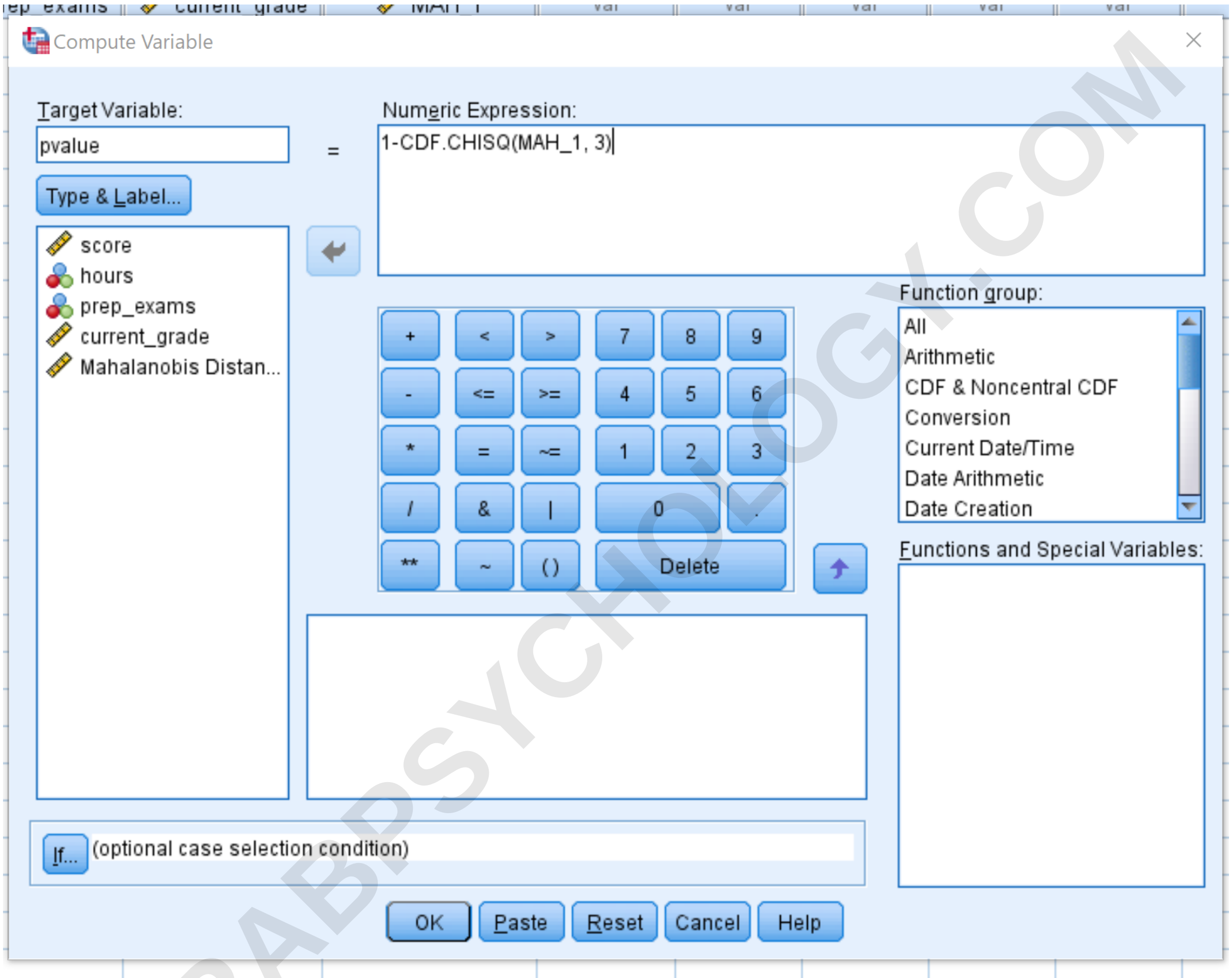
Defining the P-Value Calculation Formula

In the **Target Variable** box, assign a descriptive name for the new variable. We chose "pvalue" for clarity. Next, in the **Numeric Expression** box, input the following statistical formula, paying close attention to the degrees of freedom:

1 - CDF.CHISQ(MAH_1, 3)

The formula calculates 1 minus the cumulative probability, which yields the upper-tail probability (the p-value) corresponding to the observed Mahalanobis distance. It is critical to understand the second argument in the function: the degrees of freedom. We use **3** degrees of freedom because there were three independent variables (predictors) included in the initial linear regression model setup (Hours Studied, Prep Exams Taken, and Current Grade).

After entering the formula and confirming the correct degrees of freedom based on your specific dataset, click **OK** to run the computation. The p-values will be generated and added as a new column in the Data View.



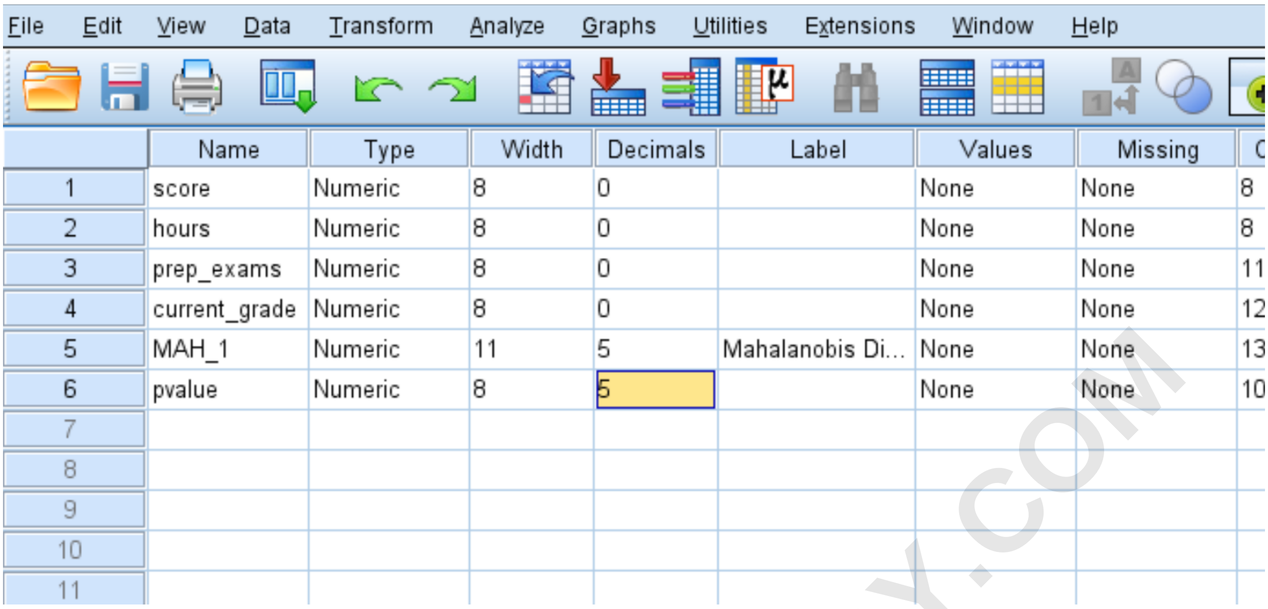
Step 4: Interpreting the P-Values and Final Outlier Identification

The new column containing the p-values will now be displayed in the Data View. This p-value is the final determinant for classifying an observation as a multivariate outlier. A smaller p-value indicates a greater statistical separation between the data point and the rest of the multivariate distribution.

score	hours	prep_exams	current_grade	MAH_1	pvalue
91	16	3	70	16.41507	.00
93	6	4	88	2.58699	.46
72	3	0	80	4.74810	.19
87	1	3	83	3.80675	.28
86	2	4	88	3.80845	.28
73	3	0	84	4.07914	.25
68	2	1	78	4.06957	.25
87	5	2	94	2.41850	.49
78	2	1	90	1.65188	.65
99	5	2	93	1.92343	.59
95	2	3	89	1.08859	.78
76	3	3	82	2.28033	.52
84	4	3	95	1.92415	.59
96	3	2	94	1.36636	.71
76	3	2	81	1.52369	.68
80	3	2	93	1.00037	.80
83	4	3	93	1.19153	.76
84	3	3	90	.66156	.88
73	4	2	89	.22777	.97
74	4	2	89	.22777	.97

Adjusting Decimal Precision for Accurate Assessment

Due to the small magnitude of significance thresholds, SPSS's default display of two decimal places is usually insufficient. To ensure rigorous assessment against the $p < .001$ threshold, navigate to the **Variable View** tab. Locate the 'pvalue' row and increase the number in the **Decimals** column to five or more decimal places.



	Name	Type	Width	Decimals	Label	Values	Missing	C
1	score	Numeric	8	0		None	None	8
2	hours	Numeric	8	0		None	None	8
3	prep_exams	Numeric	8	0		None	None	11
4	current_grade	Numeric	8	0		None	None	12
5	MAH_1	Numeric	11	5	Mahalanobis Di...	None	None	13
6	pvalue	Numeric	8	5		None	None	10
7								
8								
9								
10								
11								

Once you return to the **Data View**, the p-values will be displayed with greater precision. The statistical convention for designating a multivariate outlier is typically a p-value that is **less than .001**. This threshold ensures that we are highly confident that the data point is not part of the expected population distribution.

In our student dataset example, only the first observation meets this stringent criterion, exhibiting a p-value far below 0.001. Consequently, Observation 1 is confirmed as the only multivariate outlier in the dataset based on the Mahalanobis calculation.

	score	hours	prep_exams	current_grade	MAH_1	pvalue
1	91	16	3	70	16.41507	.00093
2	93	6	4	88	2.58699	.45977
3	72	3	0	80	4.74810	.19120
4	87	1	3	83	3.80675	.28310
5	86	2	4	88	3.80845	.28291
6	73	3	0	84	4.07914	.25304
7	68	2	1	78	4.06957	.25405
8	87	5	2	94	2.41850	.49020
9	78	2	1	90	1.65188	.64768
10	99	5	2	93	1.92343	.58845
11	95	2	3	89	1.08859	.77983
12	76	3	3	82	2.28033	.51630
13	84	4	3	95	1.92415	.58830
14	96	3	2	94	1.36636	.71344
15	76	3	2	81	1.52369	.67681
16	80	3	2	93	1.00037	.80116
17	83	4	3	93	1.19153	.75504
18	84	3	3	90	.66156	.88221
19	73	4	2	89	.22777	.97299
20	74	4	2	89	.22777	.97299
21						
22						
23						

Methodological Guidance on Handling Outliers

The identification of an outlier using the Mahalanobis distance is a critical finding, but it is not the end of the process. The presence of an outlier necessitates a careful methodological response to ensure the integrity of subsequent statistical modeling. Researchers typically have two primary options when confronted with a genuine multivariate outlier:

Verify Data Entry Accuracy

The initial response must always be a meticulous check for data entry errors. An observation with an extreme Mahalanobis distance might simply reflect a typographical mistake during data collection or transcription. If the source data can be verified and corrected, the outlier status may be resolved. If the value is confirmed to be an error but the correct value cannot be ascertained, the observation should be flagged as missing data for those specific variables, or potentially removed entirely.

Strategic Removal or Retention Decisions

If the extreme value is confirmed to be a true representation of reality--a genuine multivariate outlier--the researcher must assess its impact. If the outlier substantially alters the results of the final analysis (e.g., changes the direction or significance of key regression coefficients), the ethical and statistical choice is often to remove the data point. However, any removal must be clearly documented in the final report, justifying the methodological steps taken based on the Chi-Square distribution test of the Mahalanobis distance.

Alternatively, the researcher may perform a sensitivity analysis, comparing results derived from the full dataset versus the dataset with the outlier removed. This transparent approach demonstrates the robustness of the findings and clarifies the exact influence of the unusual observation.