

# How to Identify Outliers in SPSS: A Step-by-Step Guide

Authored by  
**stats writer**

March 15, 2026

## RECOMMENDED CITATION

stats writer (2026). *How to Identify Outliers in SPSS: A Step-by-Step Guide*.

PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=135989>

## Foundations of Outlier Detection in Statistical Research

In the realm of quantitative research, the identification of an **outlier** is a fundamental step in ensuring the integrity of any **data analysis**. An outlier is defined as an observation that deviates so significantly from other observations as to arouse suspicions that it was generated by a different mechanism. When utilizing SPSS, a leading software suite for statistical evaluation, researchers must be vigilant because these anomalies can exert an undue influence on the mean and standard deviation, potentially leading to Type I or Type II errors. Understanding the nature of these data points is critical; they may represent genuine variations in a population, or they may be the result of measurement errors or data entry mistakes that could invalidate the final conclusions of a study.

The presence of outliers often violates the assumption of **normality**, which is a prerequisite for many parametric tests, such as t-tests and ANOVA. If these extreme values are not properly identified and addressed, they can artificially inflate the variance, thereby reducing the power of statistical tests and making it more difficult to detect a significant effect where one actually exists. Therefore, the primary objective of outlier detection in SPSS is to enhance the robustness of the **statistical model**. By systematically screening the dataset, researchers can determine whether an anomalous value should be retained as a legitimate part of the sample or excluded to prevent it from skewing the results.

Effective outlier management involves a multi-faceted approach that combines **descriptive statistics** with visual diagnostics. While some researchers rely solely on mathematical thresholds, such as the Z-score, a comprehensive analysis requires a more nuanced perspective. In SPSS, the user is provided with a variety of tools to explore the distribution of data, ranging from simple frequency tables to complex multivariate distance measures. By integrating these methods, one can gain a deeper understanding of the dataset's structure and ensure that the subsequent data analysis reflects the underlying reality of the research subject rather than the noise generated by a few unrepresentative data points.

## Visualizing Data Distribution via Box Plots and Scatterplots

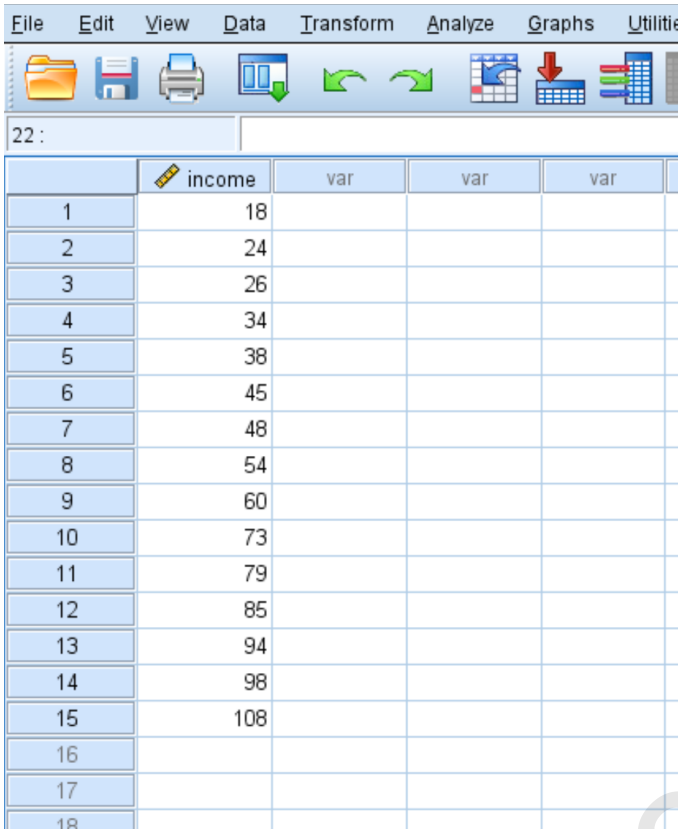
One of the most intuitive and effective ways to detect an outlier in a dataset is through visual inspection. Visual diagnostics allow the researcher to see the "shape" of the data and identify points that lie far from the central cluster. In SPSS, the box plot (also known as a box-and-whisker plot) is the gold standard for this purpose. A **boxplot** provides a graphical summary of the data's distribution based on its quartiles. It highlights the median, the spread of the middle 50% of the data, and the range of the remaining values. Any point that falls significantly outside the whiskers of the plot is flagged by the software, providing an immediate visual cue that an anomaly may be present.

Beyond the boxplot, **scatterplots** are invaluable when investigating relationships between two or more variables. While a boxplot might show an outlier in a single dimension, a scatterplot can reveal **multivariate outliers**--points that might look normal in isolation but are unusual when their relationship with another variable is considered. For instance, an annual income of \$100,000 might be typical for a professional, but if that income is associated with an age of 12 years in a dataset, it becomes a clear outlier. Visualizing these relationships helps in identifying patterns that purely numerical methods might overlook, making it a crucial first step in any exploratory data analysis within SPSS.

The visual detection phase is not merely about finding "bad" data; it is about understanding the **probability distribution** of the sample. By examining histograms and probability plots alongside boxplots, researchers can assess whether the data follows a normal distribution. If a dataset is heavily skewed, what appears to be an outlier might actually be a characteristic of a non-normal distribution, such as a log-normal or exponential distribution. Thus, visual tools in SPSS serve as a diagnostic bridge, allowing the researcher to make informed decisions about whether to apply **data transformations** or to use non-parametric statistical methods that are less sensitive to extreme values.

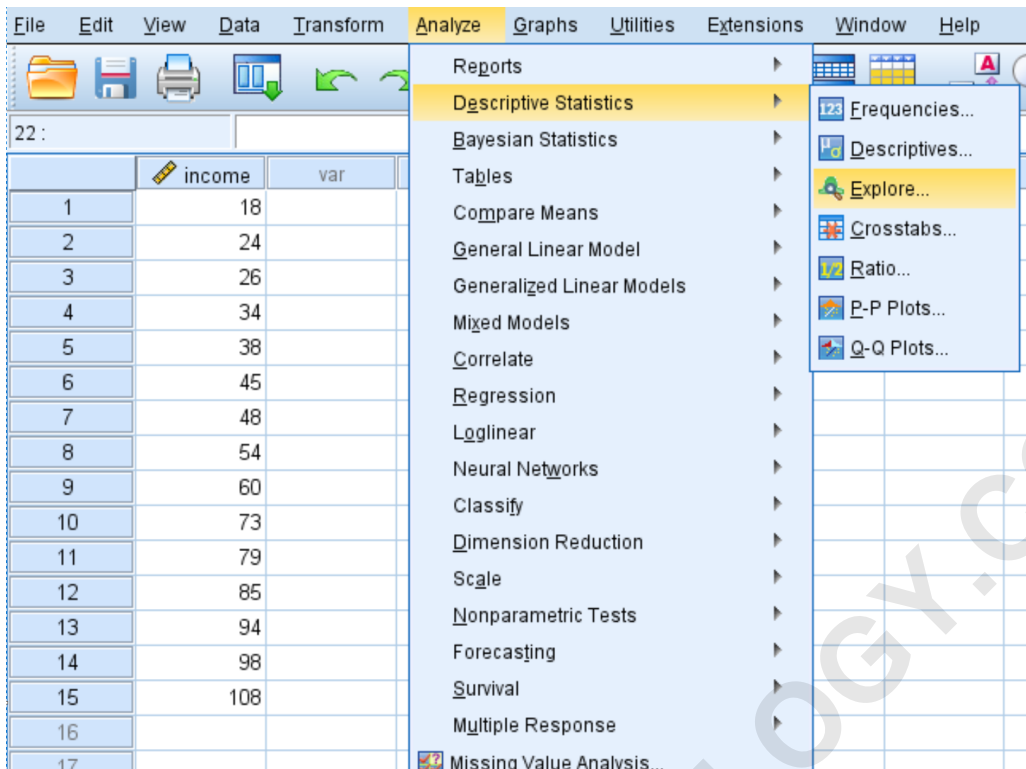
## Executing the Explore Procedure for Comprehensive Data Screening

To begin the formal process of identifying an outlier in SPSS, researchers typically utilize the "Explore" command, which is housed under the **Descriptive Statistics** menu. This procedure is designed to provide a comprehensive overview of the variables, offering both numerical summaries and visual plots. For example, consider a dataset representing the annual income of 15 individuals. To determine if any of these individuals have incomes that are uncharacteristically high or low, the researcher follows a specific sequence of commands within the software interface.

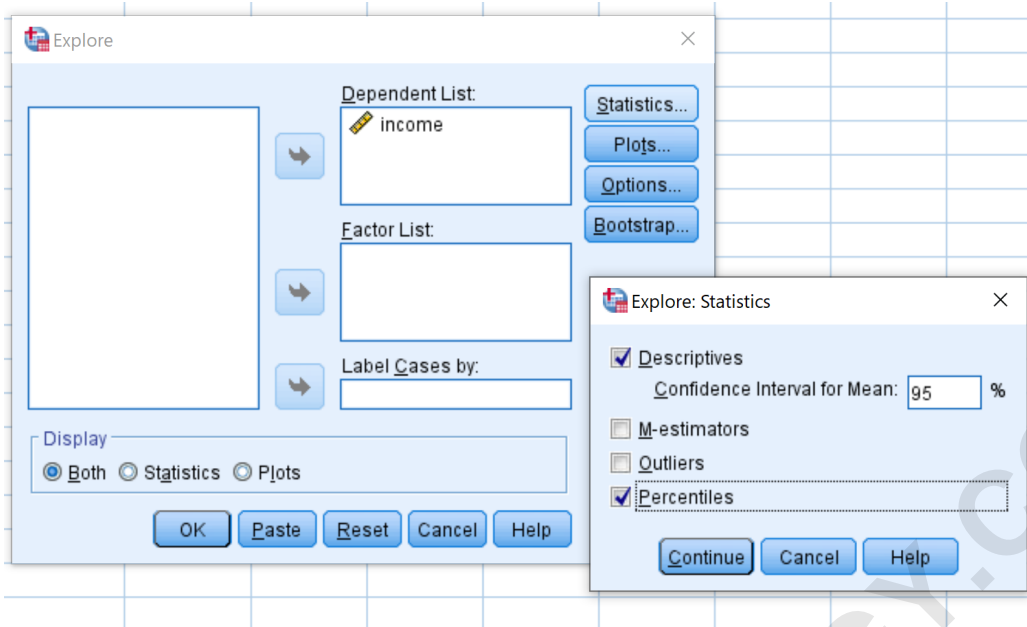


	income	var	var	var
1	18			
2	24			
3	26			
4	34			
5	38			
6	45			
7	48			
8	54			
9	60			
10	73			
11	79			
12	85			
13	94			
14	98			
15	108			
16				
17				
18				

The technical workflow involves navigating to the **Analyze** tab, selecting **Descriptive Statistics**, and then clicking on **Explore**. This opens a dialogue box where the variable of interest, such as "income," is moved into the "Dependent List." By selecting specific options within this menu, the researcher can request percentiles and descriptive measures that are essential for calculating the thresholds of an outlier. The "Explore" function is particularly powerful because it allows for the analysis of data subgroups, enabling researchers to identify outliers within specific categories or demographics simultaneously.

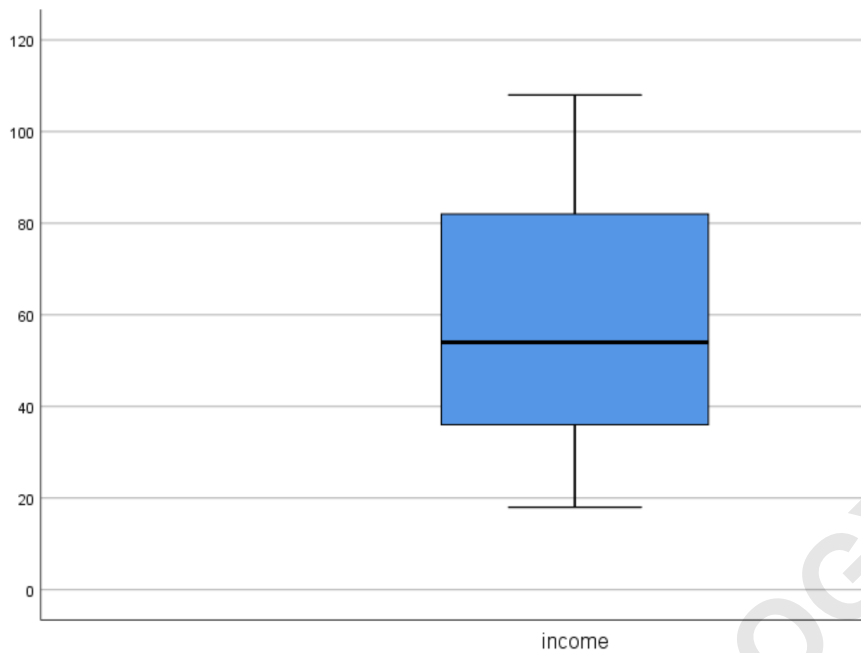


Within the "Explore" window, it is vital to click the **Statistics** button to ensure that the box for **Percentiles** is checked. This action instructs SPSS to generate the Tukey's Hinges, which are used to define the interquartile range (IQR). Once these settings are configured and the user clicks **Continue** followed by **OK**, the software processes the data and generates an output viewer containing the requested tables and the box plot. This output serves as the primary evidence for the presence or absence of anomalous data points in the sample.



The resulting box plot is a critical diagnostic tool. If the distribution is clean, the plot will show a central box with whiskers extending to the minimum and maximum values within a calculated range. However, if an outlier is detected, it will appear as a distinct symbol outside the whiskers. This automated flagging system saves the researcher from manual calculations and provides a clear, standardized method for outlier identification that is recognized across academic and professional disciplines.

**income**



**Interpreting Tukey's Hinges and the Interquartile Range**

The mathematical foundation for outlier detection in SPSS boxplots is based on the **interquartile range** (IQR). The interquartile range is defined as the difference between the third quartile (75th percentile) and the first quartile (25th percentile). In SPSS, these quartiles are often referred to as Tukey's Hinges, named after the statistician John Tukey who pioneered this method of exploratory data analysis. The IQR is a robust measure of variability because it is not influenced by extreme values, unlike the standard deviation.

		Percentiles						
		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	income	18.00	21.60	34.00	54.00	85.00	102.00	
Tukey's Hinges	income			36.00	54.00	82.00		

To determine what constitutes an outlier, SPSS applies a specific formula. Any data point is considered a "mild" outlier if it falls outside the following boundaries:

**Upper Bound:** 3rd Quartile + (1.5 \* IQR)

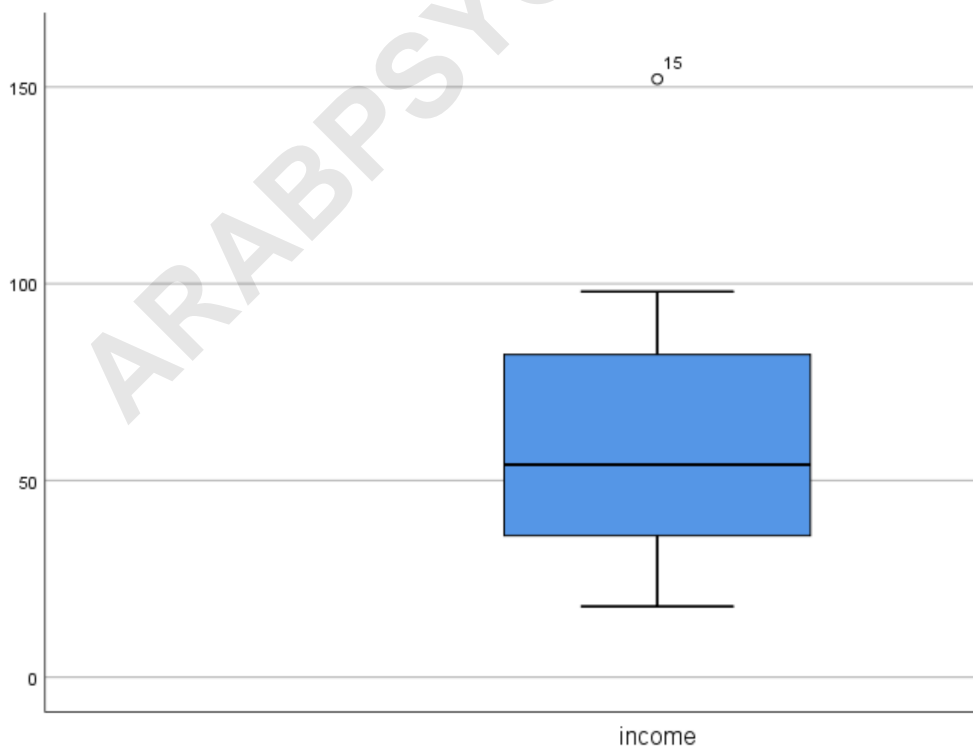
**Lower Bound:** 1st Quartile - (1.5 \* IQR)

For instance, if the 75th percentile (3rd quartile) is 82 and the 25th percentile (1st quartile) is 36, the interquartile range is calculated as  $82 - 36 = 46$ . Applying the multiplier, we find that  $1.5 * 46 = 69$ . Therefore, any value above 151 ( $82 + 69$ ) or below -33 ( $36 - 69$ ) would be flagged as an outlier. In many real-world scenarios, such as income, negative values are impossible, so only the upper bound is relevant for the analysis.

The importance of the 1.5 multiplier cannot be overstated; it represents a balance between sensitivity and specificity. While it is sensitive enough to catch values that are truly unrepresentative of the majority, it is not so strict that it flags common variations in a **normally distributed** dataset. By using this standardized mathematical approach, researchers can move away from subjective "eyeballing" of data and toward a more rigorous, defensible standard of **statistical significance**. This ensures that the identification of an outlier is consistent across different studies and researchers.

When SPSS identifies a value that exceeds these thresholds, it labels the point on the box plot with the corresponding case number from the **Data View** tab. This feature is exceptionally useful for data cleaning, as it allows the researcher to return to the specific row in the dataset and investigate the raw data. For example, if case number 15 is identified as an outlier with an income of 152, the researcher can check the original survey or data source to ensure this was not a clerical error.

### income



## Identifying Extreme Outliers through Visual Markers

In addition to mild outliers, SPSS distinguishes a more severe category known as an **extreme outlier**. While a mild outlier might be a slightly unusual observation, an extreme outlier represents a data point that is drastically far from the rest of the distribution. These points are particularly dangerous to the validity of **statistical tests** because their leverage on the results is significantly higher. In an SPSS boxplot, mild outliers are denoted by a circle (O), whereas extreme outliers are marked with an asterisk (\*).

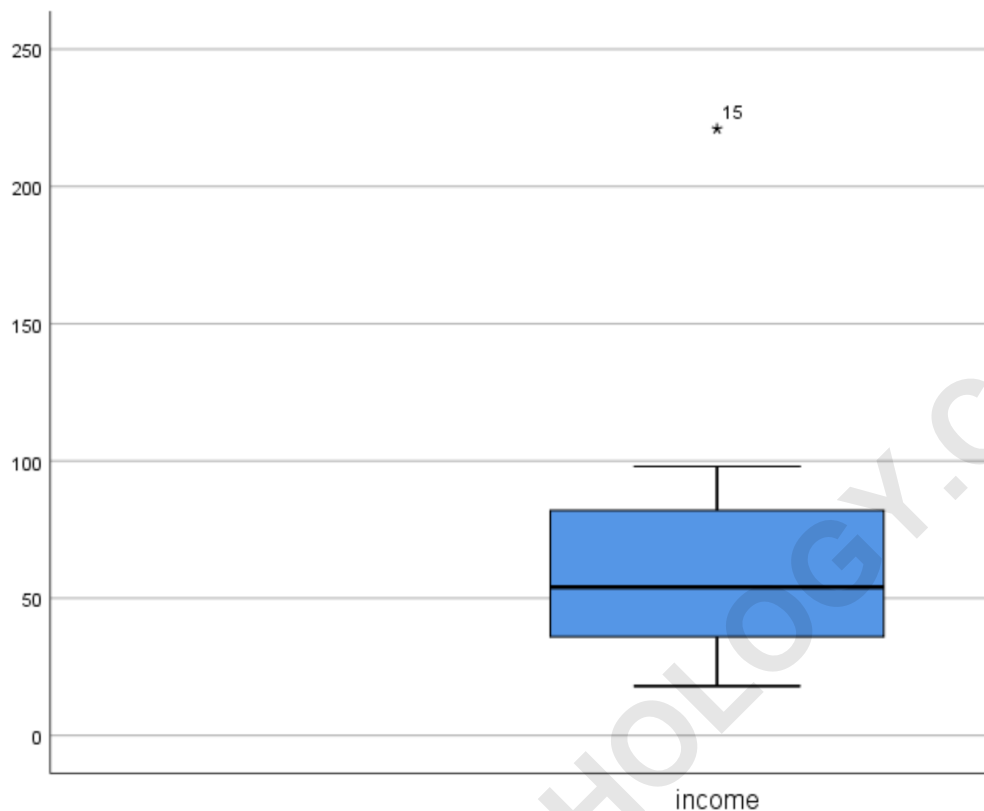
The mathematical threshold for an **extreme outlier** is much higher than that of a mild one. SPSS uses a factor of 3 times the interquartile range to define these boundaries:

**Extreme Upper Bound:** 3rd Quartile + (3 \* IQR)

**Extreme Lower Bound:** 1st Quartile - (3 \* IQR)

Continuing with the previous example where the IQR was 46, the calculation for extreme outliers would be  $3 * 46 = 138$ . Consequently, any income value exceeding 220 ( $82 + 138$ ) or falling below -102 ( $36 - 138$ ) would be classified as extreme. When such a point appears on the plot with an asterisk, it serves as a high-priority warning to the researcher that this specific case could fundamentally alter the outcome of the study.

## income



The distinction between circles and asterisks is vital for **data cleaning**. A few mild outliers might be tolerable in a large sample size, as the central limit theorem helps mitigate their impact. However, even a single extreme outlier in a small sample can shift the mean and lead to misleading correlations. By visually separating these two types of anomalies, SPSS allows researchers to prioritize their investigation, focusing first on the most impactful data points. This hierarchical approach to outlier detection is essential for maintaining a high standard of **research methodology**.

## Advanced Statistical Methods: Z-Scores and Mahalanobis Distance

While boxplots are excellent for univariate analysis, some datasets require more advanced techniques. One common method for identifying an outlier is the use of the Z-score. A **Z-score** measures how many standard deviations a data point is from the mean. In a standard normal distribution, approximately 99.7% of the data falls within three standard deviations. Therefore, any case with a Z-score greater than +3 or less than -3 is typically flagged as a potential outlier. SPSS can easily calculate these values using the "Descriptives" command and saving the standardized values as new variables.

For **multivariate analysis**, where the researcher is looking at multiple variables simultaneously, the Mahalanobis distance is the preferred statistical metric. This measure accounts for the **covariance** between variables, determining how far a point is from the center of a multi-dimensional cloud of data. Unlike univariate methods, Mahalanobis distance can detect an outlier that doesn't necessarily have extreme values in any single variable but has an unusual combination of values across several variables. This is particularly useful in **regression analysis** and **cluster analysis**.

In SPSS, Mahalanobis distance can be calculated through the **Linear Regression** menu. By selecting "Mahalanobis" under the "Save" options, the software generates a new column in the dataset representing the distance for each case. The researcher then compares these values against a chi-square distribution table to determine statistical significance. If the calculated distance exceeds the critical value for a given **alpha level** and degrees of freedom, the case is identified as a multivariate outlier that warrants further investigation.

Another sophisticated approach involves cluster analysis, which groups similar observations together. Data points that do not fit into any cluster or form very small, isolated clusters can be identified as outliers. This method is highly effective for large, complex datasets where the relationships between variables are non-linear or poorly understood. By combining these advanced statistical methods with the visual diagnostics of SPSS, researchers can build a robust defense against anomalous data, ensuring that their **empirical results** are both accurate and reproducible.

## Ethical and Practical Considerations for Handling Identified Outliers

Once an outlier has been identified, the researcher faces the critical decision of how to handle it. The first step is always to verify the **accuracy** of the data. This involves checking for **data entry errors**, equipment malfunctions, or participant non-compliance. If the outlier is the result of a clear error, it should be corrected if the true value can be retrieved, or removed if the correct value is unknown. This is the simplest form of data cleaning and is essential for maintaining the **integrity** of the dataset.

If the outlier is found to be a legitimate observation, the decision becomes more complex. One option is to **remove the outlier** entirely from the analysis. This is often done when the point is so extreme that it represents a population different from the one under study. However, researchers must be transparent about this process. It is considered a best practice in **scientific writing** to report the number of outliers removed and the criteria used for their exclusion. Arbitrarily removing data to achieve a desired p-value is a form of **data manipulation** and is ethically unacceptable.

An alternative to removal is **data transformation** or **Winsorization**. Transformation involves applying a mathematical function, such as a logarithm or square root, to the entire variable to reduce the impact of extreme values and bring the distribution closer to normality. Winsorization

involve replacing the extreme values with the nearest non-outlier values (e.g., replacing a value at the 99th percentile with the value of the 95th percentile). This approach allows the researcher to retain the case while minimizing its disproportionate influence on the **statistical analysis**.

Ultimately, the choice of how to handle an outlier depends on the research goals and the nature of the data. In some cases, the outliers themselves are the most interesting part of the study, representing "black swan" events or exceptional individuals. Regardless of the chosen path, the use of SPSS provides the technical framework necessary to identify these points with precision. By following a systematic and transparent protocol, researchers can ensure that their **data analysis** is both rigorous and honest, leading to more reliable contributions to the body of scientific knowledge.

ARABPSYCHOLOGY.COM