

Learning How to Perform Grubbs' Test for Outlier Detection in R

Authored by
Mohammed loot

November 8, 2025

RECOMMENDED CITATION

Mohammed loot (2025). *Learning How to Perform Grubbs' Test for Outlier Detection in R*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=13355>

Identifying [outliers](#) in a dataset is arguably one of the most crucial initial steps in any rigorous data cleaning or statistical analysis pipeline. An [outlier](#) is formally defined as an observation point that is significantly distant from other observations, often suggesting unusual variability, measurement errors, or unique phenomena not representative of the underlying process. If left untreated, these extreme values possess the power to severely skew statistical results, invalidate parametric model assumptions, and lead to erroneous conclusions. Consequently, analysts require a robust, statistically sound method for determining if a suspicious data point genuinely constitutes an anomaly worth investigating further or removal.

To formally test for the presence of these unusual values, statisticians frequently employ [Grubbs' Test](#), which is also widely recognized as the Maximum Normed Residual Test. This powerful statistical procedure is specifically designed to determine if the maximum or minimum value in a sample drawn from a [normally distributed](#) population is statistically different from the rest of the data points. The successful application of [Grubbs' Test](#) provides empirical evidence to support or reject the notion that an extreme data point should be classified as a true outlier.

The Critical Role of Outlier Detection in Data Analysis

While visual inspection, such as box plots or scatter plots, can hint at the presence of anomalies, relying solely on graphical methods introduces subjectivity and lacks statistical rigor. [Grubbs' Test](#) shifts the determination from visual estimation to formalized [hypothesis testing](#), lending statistical credibility to the data cleaning process. This test is particularly important in fields requiring high precision, such as quality control, financial modeling, or physical sciences, where a single measurement error could drastically alter interpretations. By calculating the difference between the suspected outlier and the sample mean, relative to the standard deviation, the test provides a quantifiable measure of unusualness.

It is essential to understand that not all extreme values are true outliers that should be removed. Some merely represent the natural variability of the system under observation. However, if an extreme value results in a rejection of the null hypothesis in [Grubbs' Test](#), it indicates a high probability that the observation does not belong to the same population distribution as the remaining data. This necessitates careful investigation into the data collection methodology or the underlying physical process that generated the data point. This tutorial focuses on executing this vital procedure within the powerful statistical programming environment, [R](#).

Prerequisites and Environment Setup for Grubbs' Test in R

Before attempting to execute Grubbs' Test, analysts must rigorously confirm that their data meets two foundational statistical prerequisites to ensure the validity and reliability of the output. The most critical assumption is that the dataset must be approximately [normally distributed](#). Grubbs' Test

is a parametric test, meaning it relies heavily on the data conforming to a normal distribution; failure to meet this condition may lead to incorrect inference regarding the presence of anomalies. Secondly, the test is sensitive to sample size, requiring a minimum of seven observations ($N \geq 7$). If the sample size is too small, the test lacks the statistical power necessary to make reliable judgments about deviation.

Once these data prerequisites are satisfied, the **R** environment must be configured correctly. The functionality for performing Grubbs' Test is not included in the base **R** installation; rather, it resides within the community-maintained **Outliers** package. This package is specifically designed to provide a suite of statistical tests for identifying various types of anomalies. Therefore, the essential first step involves installing this package from CRAN and subsequently loading it into the current session using the standard `install.packages()` and `library()` commands.

Deep Dive into the `grubbs.test()` Function and Core Hypotheses

The primary function utilized for this analysis is `grubbs.test()`, which is provided by the **Outliers** package. Mastering the syntax and understanding the parameters of this function is essential, as they allow sophisticated control over which type of **outlier** detection is performed--whether testing the highest value, the lowest value, or even testing for multiple anomalies simultaneously. The standard function signature provides the necessary flexibility for comprehensive analysis in virtually any data scenario.

The general structure of the function call is defined as follows:

`grubbs.test(x, type = 10, opposite = FALSE, two.sided = FALSE)`

Each argument plays a distinct and crucial role in defining the scope and nature of the hypothesis test being performed. Understanding these parameters ensures that the correct form of the test is applied to match the analyst's specific concerns regarding the dataset:

x: This is the mandatory input parameter, representing the numeric vector containing the data values that are to be tested for the presence of **outliers**.

type: This argument specifies the number and location of the potential anomalies being assessed. The default setting is designed for the most common use case:

type = 10 (Default): Executes a test to determine if the single maximum value in the sample is an outlier.

type = 11: Tests simultaneously if both the minimum and maximum values are single outliers (this setting often necessitates a two-sided test).

type = 20: Used when testing for the presence of two outliers situated on the same tail (i.e., testing if the two highest values or the two lowest values are both anomalous).

opposite: A logical value (TRUE/FALSE) used exclusively when testing for a single outlier (i.e., when `type` is 10). If set to **TRUE**, the test pivots its focus from the maximum value (the default behavior) to the minimum value, effectively checking if the lowest observation is the anomaly.

two.sided: A logical value indicating whether the test should be performed as a two-sided test, which checks for outliers on both the high and low ends of the distribution simultaneously, or as a one-sided test (the default).

In the context of formal statistical testing, establishing the competing hypotheses is paramount before running the analysis. Grubbs' Test operates under the following framework:

H0 (null hypothesis): There is no statistical outlier present in the provided data sample. All observations belong to the same normally distributed population.

HA (alternative hypothesis): There is at least one statistical outlier present in the data sample. The extreme observation(s) do not belong to the same normally distributed population.

Case Study 1: Identifying a Single Maximum Outlier

To illustrate the implementation of the default Grubbs' Test, consider a realistic scenario where we have collected 18 observations and suspect that the highest recorded value might be statistically anomalous. We will utilize the default settings (**type=10** and **opposite=FALSE**), which specifically targets the observation most distant from the mean in the positive direction.

Our sample dataset ranges from 5 to 40. The maximum value of 40 appears significantly larger than its nearest neighbors, making it the primary candidate for rejection as a representative data point. We begin by loading the necessary **Outliers** package and defining our data vector within the **R** environment before executing the primary test function. This structured approach ensures reproducibility and clear documentation of the steps taken.

```
#load Outliers package
```

```
library(Outliers)
```

```
#create data
```

```
data <- c(5, 14, 15, 15, 14, 13, 19, 17, 16, 20, 22, 8, 21, 28, 11, 9, 29, 40)
```

```
#perform Grubbs' Test to see if '40' is an outlier
```

```
grubbs.test(data)
```

```
# Grubbs test for one outlier
```

```
#
```

```
#data: data
```

```
#G = 2.65990, U = 0.55935, p-value = 0.02398
```

```
#alternative hypothesis: highest value 40 is an outlier
```

The execution yields critical statistical metrics. We are presented with the test statistic, designated as **G = 2.65990**, and its corresponding **p-value = 0.02398**. In standard statistical practice, a significance level (alpha, α) is typically set at 0.05. Since our calculated **p-value** (0.02398) is decisively less than this threshold (0.05), we possess sufficient statistical evidence to reject the **null hypothesis** (H_0). Our conclusion, therefore, aligns with the alternative hypothesis: the maximum value of 40 is indeed classified as a statistically significant outlier within this specific dataset.

Case Study 2: Testing Minimum Values and Multiple Anomalies

One of the greatest strengths of the **grubbs.test()** function is its versatility, allowing us to pivot the focus of the test instantly. If we were concerned about the lowest value in our original dataset (which is 5) rather than the confirmed maximum outlier (40), we must perform an independent check. To shift the test's attention from the upper tail to the lower tail, we simply utilize the **opposite=TRUE** argument within the function call, while keeping the default single-outlier type (`type=10`).

```
#perform Grubbs' Test to see if '5' is an outlier
grubbs.test(data, opposite=TRUE)
```

```
# Grubbs test for one outlier
#
#data: data
#G = 1.4879, U = 0.8621, p-value = 1
#alternative hypothesis: lowest value 5 is an outlier
```

For this second iteration, the test statistic is calculated as **G = 1.4879**, resulting in a **p-value** of 1. Given that this **p-value** is substantially greater than the conventional 0.05 significance threshold, we lack the necessary statistical evidence to reject the **null hypothesis**. Consequently, we conclude that the minimum value of 5, while the smallest observation, does not qualify as a statistically significant outlier relative to the distribution of the remaining sample data.

Furthermore, **grubbs.test()** can be adapted to detect multiple outliers within the same tail. Consider a modified dataset incorporating two large, potentially anomalous values: 40 and 42. To test if **both** of these values should be rejected simultaneously, we must specify the argument **type=20**. Note that detecting two outliers generally requires a slightly larger minimum sample size ($N \geq 11$) to maintain the test's statistical power and validity, though our current sample size is adequate.

```
#create dataset with two large values at one end: 40 and 42
data <- c(5, 14, 15, 15, 14, 13, 19, 17, 16, 20, 22, 8, 21, 28, 11, 9, 29, 40, 42)

#perform Grubbs' Test to see if both 40 and 42 are outliers
grubbs.test(data, type=20)

# Grubbs test for two outliers
#
#data: data
#U = 0.38111, p-value = 0.01195
#alternative hypothesis: highest values 40 , 42 are outliers
```

For this two-outlier investigation, the resulting **p-value** is calculated as **0.01195**. Since this value is significantly below the 0.05 threshold, we confidently reject the **null hypothesis**. This result provides robust statistical evidence supporting the conclusion that the values 40 and 42 are simultaneously classified as outliers in this modified dataset, warranting further scrutiny or treatment.

Methodological Decisions: What to Do After Identifying an Outlier

Successfully identifying an anomaly using Grubbs' Test marks the transition from detection to mitigation. The presence of a statistically extreme value mandates a careful methodological decision, as the subsequent choice--whether to retain, modify, or remove the data point--will profoundly influence the results of any downstream analysis and model fitting. Data scientists typically approach the management of detected outliers by considering three primary courses of action, prioritized based on investigation and data integrity:

1. Verification of Errors and Typos. The most critical initial step is thorough verification. Often, values flagged as **outliers** are not reflections of true phenomena but are artifacts of human or instrumentation error. This includes simple transcription mistakes, misplaced decimal points, or incorrect unit entries during data collection or input. Analysts must mandate a comprehensive review of the raw data source and the entire data entry process. If the extreme value is confirmed to be an error, it should be corrected to its true value or, if correction is impossible, treated as missing data rather than immediately subjected to removal based solely on its statistical distance.

2. Assignment of a New Value (Imputation or Winsorizing). If the outlier is determined to be the result of a known, uncorrectable measurement error, or if the original value cannot be definitively verified, analysts might opt to assign a new, more representative value. This process, known as **imputation**, often involves replacing the outlier with a measure of central tendency for the dataset, such as the mean or median. A more sophisticated alternative is Winsorizing, where the outlier

value is replaced by the nearest non-outlying observation, effectively capping the extreme value without removing the observation entirely. This method preserves the sample size while reducing the disproportionate influence of the anomaly.

3. Removal of the Outlier. The decision to entirely remove an observation should be treated as a last resort and must be exercised with extreme caution and complete transparency. If the extreme value is confirmed to be a true observation (not an error) but its existence fundamentally violates critical assumptions of the intended statistical model (e.g., severe non-normality caused by a single point), removal may be considered. Removal is generally only advisable when the underlying mechanism generating the outlier is fundamentally different from the process being studied (e.g., a technical malfunction vs. natural variation) or if the analysis explicitly requires a dataset free of such influential points. Documentation of the removal criteria is mandatory for ethical statistical reporting.