

Understanding Outliers and Their Effect on Calculating the Mean

Authored by
Mohammed looti

November 8, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Understanding Outliers and Their Effect on Calculating the Mean*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=13928>

Defining the Arithmetic Mean and Its Role in Descriptive Statistics

The [arithmetic mean](#), conventionally referred to simply as the **mean** or average, stands as the most fundamental measure of [central tendency](#) within the field of [statistics](#). Its primary function is to represent the typical or expected value within a specific dataset, thereby offering a crucial summary of where the numerical "center" of the data distribution resides. This single, easily interpretable figure is essential for descriptive statistics, enabling analysts to condense extensive data into a manageable summary statistic. A solid comprehension of the mean is a prerequisite for advanced statistical procedures, including [hypothesis testing](#) and advanced regression modeling.

The calculation of the mean is deceptively simple: one sums all observed values in the dataset and subsequently divides this total by the number of observations. This straightforward methodology accounts for its ubiquitous application across diverse disciplines, spanning from economic forecasting and financial risk assessment to rigorous scientific experimentation. However, despite its computational ease, the interpretation of the mean necessitates careful consideration, especially when analysts are dealing with raw, potentially contaminated datasets.

To illustrate the foundational calculation, consider the mathematical definition:

Mean = (Sum of all observations) / (Total count of observations)

For a small, controlled dataset, the calculation results in $(1 + 4 + 5 + 6 + 7) / 5$, yielding a **mean** of 4.6. While this value effectively serves as the representative midpoint for this balanced collection, the utility of the mean is significantly compromised by its inherent sensitivity to extreme values. These extreme values, critically known as **outliers**, can dramatically shift the calculated mean, meaning the statistic may not accurately reflect the true central location of the majority of the data points, particularly if the distribution exhibits pronounced [skewness](#).

The Definition and Impact of Statistical Outliers

An [outlier](#) is formally defined as an observation that deviates significantly from other observations in a data sample. Essentially, outliers are values that fall outside the expected range or pattern exhibited by the bulk of the data. The origins of these extreme points are varied; they might stem from simple measurement inconsistencies, fundamental errors during data entry, or they may represent genuine, albeit rare, variability inherent to the population under study. Recognizing and meticulously investigating the underlying cause of these extreme values constitutes a vital step in ensuring robust and reliable data analysis.

The presence of even one or two outliers exerts a disproportionate leverage over the calculation of the mean, invariably dragging it toward the direction of the extreme value. This influence is especially pronounced when the overall [sample size](#) of the dataset is limited, as the statistical

weight of the outlier becomes magnified relative to the smaller total number of observations. Consequently, although the mean is intended to summarize the **central tendency**, it can rapidly become a misleading statistic if the analyst fails to appropriately account for the influence of these distorting outliers.

A classic anecdotal illustration, frequently cited in introductory statistical courses, dramatically demonstrates how an outlier can severely skew the calculated average:

Imagine ten people seated in a room, whose collective average income is \$50,000. If one person exits and is immediately replaced by a high-net-worth individual, such as Bill Gates, the new calculated average income for the ten individuals in the room might immediately jump to \$40 million.

This hyperbolic scenario starkly illustrates the core principle: a single, extraordinarily large value pulls the calculated mean far away from the typical income shared by the remaining nine individuals. The resulting mean of \$40 million, while mathematically sound, utterly fails to represent the financial standing of the vast majority of people in that specific location.

Directional Distortion: How Outlier Magnitude Affects the Mean

The distortion introduced by an outlier is not restricted solely to unusually high values; the mean is equally vulnerable to the influence of observations that are significantly smaller than the rest of the dataset. The effect is directional: an exceptionally large outlier induces positive [skewness](#), pulling the mean upwards, whereas an unusually small outlier causes negative skewness, dragging the mean downwards. In either instance, the mean loses its effectiveness as a reliable indicator of the distribution's center, thereby presenting a skewed and often inaccurate picture of the dataset.

The scenario involving the high-income individual demonstrated a case of positive skewness. Conversely, observations that are unusually small are common in performance measurements, such as standardized test scores, where a score of zero or near-zero might represent non-participation, technical failure, or a significant deviation from the cohort's performance norm.

Consider a practical scenario involving student exam performance to quantify the effect of a small outlier:

Ten students take an examination and receive the following scores:

The calculated mean score for this group is **84.6**.

However, if the single score of "0" is removed from the dataset (perhaps identified as an anomaly or non-participation), the mean score for the remaining nine students increases sharply to **94**.

The resulting difference between 84.6 and 94 is substantial, unequivocally demonstrating how one unusually low score can significantly depress the average performance for the entire cohort. This single point masks the fact that the typical performance level was actually situated in the mid-nineties. This example underscores the necessity for data analysts to meticulously investigate all extreme values before relying exclusively on the mean for any interpretive conclusions.

The Inverse Relationship Between Sample Size and Outlier Influence

The degree to which an outlier can distort the mean is directly dependent upon the [sample size](#) of the underlying dataset. This relationship is fundamentally inverse: in a small dataset, each individual data point contributes a larger proportion to the overall total sum, thereby amplifying the potential impact of any single extreme value. Conversely, as the sample size expands--moving towards the realm of big data--the influence exerted by a single outlier diminishes significantly because its contribution relative to the aggregate sum becomes statistically negligible.

This principle is foundational to assessing the robustness of statistical measures. If an analyst works with a small sample (e.g., $n=10$), a single outlier will drastically skew the results, potentially rendering the mean an inaccurate descriptor of the population. However, if the sample size is extensive (e.g., $n=1000$), the identical outlier might cause only a modest and statistically insignificant shift in the calculated mean.

To further clarify this concept, let us revisit the student exam score scenario, but now assume a substantially larger sample size. Suppose a dataset contains 100 exam scores, where 99 students scored 90 or higher, and only one student recorded a score of zero. The zero remains an extreme outlier, yet its influence is buffered by the large volume of consistent data points:

The mean of the 100 scores is calculated to be **93.18**. If the score of "0" were removed, the mean of the remaining 99 scores would be **94.12**.

The resulting difference in means (0.94) is relatively minor when compared to the nearly 10-point shift observed in the small sample size example. This comparison clearly highlights that even an extreme outlier has only a modest effect on the mean when the dataset is sufficiently large, granting a crucial degree of inherent stability to analyses conducted on expansive data collections.

Systematic Strategies for Handling Influential Outliers

When preparing data for analysis, encountering an outlier requires a thoughtful, strategic decision regarding its appropriate treatment. While ignoring potentially influential outliers can lead to flawed conclusions, the arbitrary removal of data points can introduce bias and reduce the generalizability of the findings. Effective [data cleaning](#) protocols mandate systematic procedures to ensure that the final processed dataset accurately represents the underlying phenomenon being investigated.

Analysts typically employ a variety of methodological options to mitigate the distorting effects of outliers on the mean. These strategies range from simple verification and correction to more complex statistical transformation or substitution methods, dependent primarily on the identified nature and source of the extreme value. The overriding objective must always be to maintain data integrity while simultaneously achieving a reliable measure of **central tendency**.

If analysis indicates the presence of an outlier unduly influencing statistical results, the following methodological options should be systematically considered:

Validate the Source and Entry Accuracy: The foundational step involves rigorously checking that the outlier is not simply the product of a basic **data entry error** or a malfunction in the measurement apparatus. For example, a value incorrectly logged as 1,000 instead of 100 is a processing artifact that must be corrected. If verification confirms an input error, the value should be corrected to the true observation, or if the true value cannot be recovered, it should be flagged as missing data.

Assign a Substitute Value (Imputation or Trimming): If the outlier is confirmed to be an error that cannot be corrected, or if the outlier is legitimate but its magnitude is highly influential, analysts may opt to assign a new, less extreme value. This process, often referred to as **imputation** or winsorizing, might involve substituting the outlier with the dataset's [median](#), the average of the nearest non-outlying points, or by capping the value at a predetermined percentile (e.g., the 95th or 99th percentile) to reduce its extreme influence.

Judicious Removal of the Outlier: If the extreme value is determined to be a genuine anomaly--perhaps representing a distinct population subgroup irrelevant to the current study or a unique, non-replicable event--and its impact is critically distorting the overall analysis, calculated removal may be warranted. However, this action requires complete documentation and justification in the final report, explicitly detailing the criteria for removal and including a sensitivity analysis to compare results with and without the excluded observation.

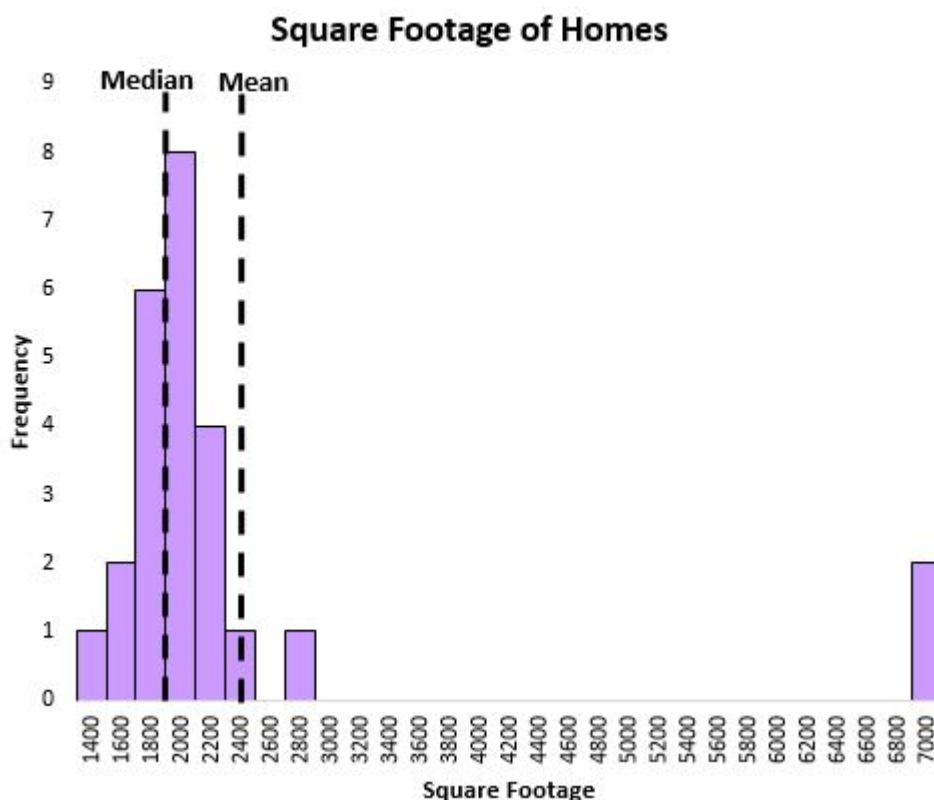
The Median: A Robust Alternative Measure of Central Tendency

Given the established vulnerability of the mean to extreme values, statisticians frequently employ the **median** as a significantly more robust alternative for describing the center of a distribution, particularly when the data is known to be asymmetrical or contaminated by influential outliers. The median is defined simply as the middle value in a dataset after all observations have been sequentially ranked in ascending or descending order. If the dataset contains an even number of observations, the median is conventionally calculated as the average of the two middle values.

The core advantage of the [median](#) is its innate resistance, or robustness, to extreme values. Unlike the mean, which integrates the actual magnitude of every single observation into its calculation, the

median relies exclusively on the rank order of the data points. Consequently, dramatically altering the magnitude of the single largest or smallest value in the dataset will not affect the median, provided that data point retains its relative position near the distribution's middle.

This statistical resilience makes the median exceptionally valuable when analyzing data concerning wealth, income distribution, or property values, domains where distributions are naturally asymmetrical and often contain substantial outliers. Consider the following visualization which compares the mean and median for housing square footage in a neighborhood:



In this housing context, the mean is heavily influenced by a few extremely large properties, erroneously inflating the average and suggesting a "typical" size far greater than what most residents possess. Conversely, the median remains centered, providing a superior and more faithful representation of the central or typical square footage found within that neighborhood. Thus, when interpreting skewed data or data potentially compromised by outliers, the median frequently serves as a superior and more informative measure of **central tendency** than the arithmetic mean.