

Missing Data Before Analysis

Knowing how many values are missing is only the first step. What matters most for your analysis is why they are missing, a question a missingness count alone cannot answer.



StatReason

Missing Data Before Analysis

Exported from statreason.com. Deterministic, browser-local content; not professional statistical or research advice.

What this answers

This page answers "I have missing values in my dataset, what do I need to know before analyzing what remains?" A count and pattern of missing values is a useful starting point, but the mechanism behind why those values are missing determines whether analyzing the remaining complete cases is safe or misleading.

Three missingness mechanisms

Data missing completely at random means the chance a value is missing has no relationship to any variable, observed or unobserved; analyzing complete cases in this situation is generally safe, though it does reduce your effective sample size. Data missing at random means the chance of missingness relates to other observed variables in your data, which some methods can account for. Data missing not at random means the chance of missingness relates to the value that would have been observed itself, such as people with the highest incomes being least likely to report income; this is the most dangerous case and complete-case analysis can produce a systematically biased result.

Why a count alone cannot tell you the mechanism

A summary of how many values are missing in each column tells you the scope of the problem but says nothing about its cause. Two datasets with identical missingness counts can have completely different mechanisms behind them, one safe to analyze with complete cases and

one not, and no automated check on the data alone can fully distinguish them without knowledge of how the data were actually collected.

Worked example

A survey asks about household income and 15 percent of respondents skip that question. If missingness is unrelated to income itself, perhaps caused by a technical form glitch affecting a random subset of respondents, the remaining 85 percent still represents the population reasonably well. If instead higher earners are more likely to skip the question out of privacy concern, the remaining respondents skew toward lower incomes, and any average computed from the complete cases will understate the true population average.

Assumption audit

Calculated from your data: the exact count and percentage of missing values per variable, and which rows have any missing value, from the linked summary engine.

Evidence to review: whether missingness in one variable correlates with other observed variables, which can suggest, though never fully confirm, the underlying mechanism.

You must verify: how the missing values actually arose in your specific data-collection process, since that context, not the data itself, is what determines which mechanism applies.

Source

This guidance follows the missing-data mechanism framework in the NIST/SEMATECH e-Handbook of Statistical Methods and the shared statistical reasoning contract every StatReason engine is built against.

Limitations

This page explains the concepts behind missing-data mechanisms; it does not cover imputation methods, which require additional modeling choices beyond the scope of this site's deterministic, browser-local calculators.

Next action: summarize your missing values with the **Missing Data Summary** before running any analysis on the remaining complete cases.