How to Perform a G-Test of Goodness of Fit to Validate Your Data

Authored by

stats writer

January 22, 2026

RECOMMENDED CITATION

stats writer (2026). *How to Perform a G-Test of Goodness of Fit to Validate Your Data*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=127079

ARABPSYCHOLOGY.COM

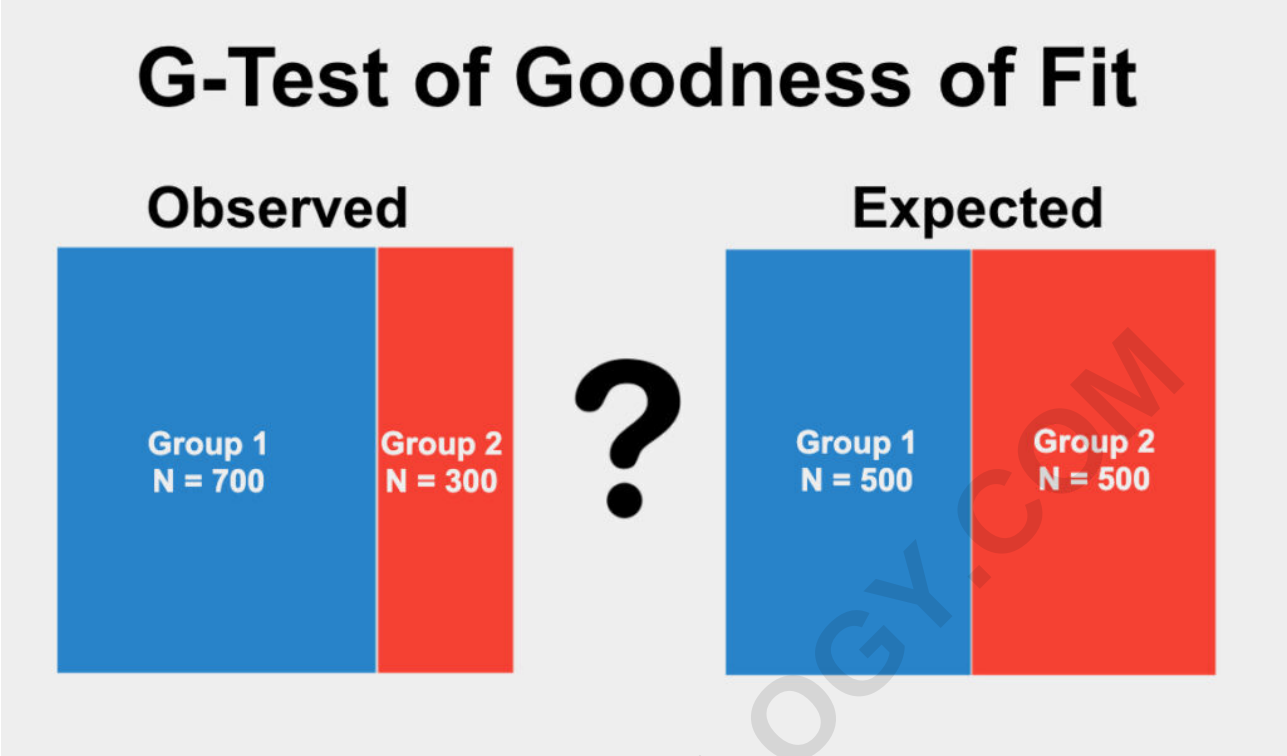
The G-Test of Goodness of Fit is a powerful statistical test employed to assess whether the frequency distribution of observed data differs significantly from a hypothesized theoretical distribution. This method is particularly suitable for analyzing categorical data, allowing researchers to compare measured outcomes against proportions predicted by a null hypothesis. The core mechanism of the G-test involves calculating the ratio of observed frequencies to expected frequencies (often using the natural logarithm), producing a G-statistic. This statistic helps quantify the discrepancies between the data and the model, enabling the determination of whether these differences are statistically significant. Consequently, the G-Test serves as an invaluable diagnostic tool across diverse fields, including **population genetics**, **ecology**, and **A/B testing**, offering clear insights into underlying data patterns and validating theoretical models.

What is the G-Test of Goodness of Fit?

The G-Test, often referred to fully as the **G-Test of Goodness of Fit**, is fundamentally designed to evaluate how well an observed set of frequencies aligns with a set of frequencies expected under a specific theoretical model or hypothesis. In practical application, this test is used when you have a single qualitative variable and you need to ascertain if the proportions observed across its categories significantly deviate from what is anticipated based on prior knowledge or theoretical prediction. For instance, if a manufacturer claims that 70% of a certain product will be defect-free, the G-test can analyze a sample batch to see if the observed defect rate is consistent with that 70% expectation.

While its primary use is comparing observed proportions to a known population proportion, the G-Test is also highly effective for comparing frequency distributions across multiple independent samples. It is mathematically similar to Pearson's Chi-Square Test, but the G-Test relies on the log-likelihood ratio, often providing a more statistically robust result, particularly when dealing with large sample sizes. It is crucial to remember the stringent sample size requirements for this asymptotic test: generally requiring a high number of observations overall (often exceeding 1000) and sufficient counts within each cell (typically more than 10) to ensure the accuracy of the resulting G-statistic and its associated p-value.

ARABPSYCHOLOGY.COM



*The G-Test of Goodness of Fit is also recognized by several other names, most notably the **G-Test**, the **Likelihood Ratio Test**, or sometimes the **Log-Likelihood Ratio Test**, reflecting its dependence on logarithmic calculations of probability ratios.*

Core Assumptions for Applying the G-Test of Goodness of Fit

Like all inferential statistical tests, the G-Test of Goodness of Fit relies on several core assumptions about the structure and collection of the data. Failing to meet these assumptions can lead to unreliable results, potentially causing incorrect rejection or acceptance of the null hypothesis. Understanding and verifying these prerequisites is a crucial step before calculating the G-statistic.

The fundamental assumptions required for accurate application of the G-Test include:

Random Sample
Sufficient Sample Size (Cell Counts)
Mutual Exclusivity of Observations

Let us delve deeper into the meaning and implications of each of these critical requirements.

The Requirement of a Random Sample

ARABPSYCHOLOGY.COM

The foundation of valid statistical inference rests upon the principle of random sampling. Specifically, every data point used in the analysis must originate from a simple random sample of the population of interest. This means that if the goal is to compare the gender ratio in a specific community sample to the national population figures, the process used to select individuals for the sample must ensure that every member of the target population had an equal and independent chance of being included.

If the selection process is non-random, the resulting sample is prone to statistical bias--a systematic error that favors certain outcomes over others. Such bias invalidates the core assumption that the sample is representative of the true population distribution, thereby rendering the results of the G-Test inaccurate and misleading, regardless of the calculated G-statistic value or p-value.

Adequate Sample Size and Expected Cell Counts

The G-Test is an asymptotic test, meaning its statistical properties (specifically, the fact that the G-statistic approximates the Chi-Square distribution) rely heavily on having a sufficient volume of data. The rule of thumb for sample size adequacy dictates that the **expected frequency** for every single category, or "cell," in your comparison must be large enough. A frequently recommended minimum is that each expected cell count should be greater than or equal to 10 subjects or participants.

This constraint is vital because low expected cell counts can severely distort the approximation of the Chi-Square distribution, leading to inflated Type I error rates (false positives). If, for example, a study categorizes participants by four income levels, each of those four income categories must be expected to contain at least 10 observations based on the population proportions being tested.

Requirement for Mutually Exclusive Observations

The requirement for mutual exclusivity ensures that the observations are independent and that each subject contributes to only one category of the variable under investigation. In the context of the G-Test of Goodness of Fit, this means that every single unit of observation--whether it is an individual survey response, a biological sample, or a transactional record--must be assigned to one and only one group or condition defined by the categorical variable. This avoids double counting and ensures the observed frequencies accurately reflect distinct events.

For instance, if a researcher is classifying types of injuries (e.g., sprain, fracture, abrasion), a single patient observation must be placed exclusively into one category. Violating the assumption of mutual exclusivity fundamentally breaches the principle of independence, which is central to most statistical tests, leading to invalid frequency comparisons.

ARABPSYCHOLOGY.COM

Deciding When to Apply the G-Test of Goodness of Fit

The G-Test of Goodness of Fit is appropriate only when your research question and data structure align with several specific criteria. These criteria help differentiate the G-Test from other inferential statistical methods, ensuring that you choose the most powerful and appropriate analysis for your dataset. Identifying the correct test type is often the most challenging part of statistical analysis.

You should strongly consider using the G-Test when all of the following conditions are met:

You are testing for a **Difference** or disparity in proportions.
Your primary variable of interest is inherently **Proportional or Categorical**.
The variable contains **Two or more options (categories)**.
The dataset satisfies the requirement of **Sufficient Observations** (typically >10 in each cell and >1000 observations overall).

Understanding the nuanced meaning of these points is key to correctly applying the G-Test.

Focusing on Differences vs. Relationships or Predictions

The G-Test is explicitly designed to test for a **difference** between observed frequencies and a set of expected frequencies, typically derived from a null hypothesis stating no difference exists (e.g., equal proportions, or proportions matching a historical baseline). It is essential that your research goal is focused on finding disparities in distribution rather than exploring relationships or predicting outcomes.

For instance, if you are analyzing election data, the G-Test answers: "Does the proportion of votes received by Candidate A significantly differ from the pre-election poll prediction of 40%?" This contrasts with tests focused on association (e.g., is gender related to voting choice?) or prediction (e.g., can age predict the likelihood of voting?).

Data Must be Proportional or Categorical

The input data for the G-Test must be counts or frequencies corresponding to a **categorical variable**. A categorical variable, also known as a qualitative variable, is one where observations fall into discrete, non-ordered groups. Classic examples include **eye color** (blue, brown, green), **type of transportation** (bus, car, bike), or **outcome of an experiment** (success, failure, inconclusive).

Proportional variables are fundamentally derived from these counts, representing the fraction or percentage of observations within each category. Examples include conversion rates (12% vs 15%), survival rates in medical trials, or the proportion of citizens who participated in a referendum.

ARABPSYCHOLOGY.COM

If your variable is **continuous** (e.g., height, temperature, income measured in dollars), the G-Test is inappropriate. Instead, if you sought to compare a continuous sample mean against an expected population mean, you might utilize a Single Sample Z-Test or a One-Sample T-Test.

If you have a continuous variable that you want to compare to an expected population mean, you should consider using a Single Sample Z-Test or a similar parametric technique.

The Categorical Variable Must Have Multiple Options

For the G-Test of Goodness of Fit to be calculated, the single categorical variable under examination must possess at least two possible outcomes or **options**. This minimum requirement allows for a comparison of observed counts against expected counts, thereby generating the necessary degrees of freedom for the test statistic. Variables with exactly two options are known as binary or dichotomous variables (e.g., recovered from illness: yes/no, or coin flip outcome: heads/tails).

However, the test is also perfectly suited for multinomial variables with many categories, such as political party affiliation (Republican, Democrat, Independent, Other) or product feedback ratings (Excellent, Good, Fair, Poor). If your variable is dichotomous (only two options) and your cell counts are very small (typically fewer than 10), the G-Test's asymptotic assumption fails, and alternative exact tests are required.

*If you possess only two options and the cell count for one or both options falls below the minimum threshold (e.g., fewer than 10 observations), it is advisable to use the **Binomial Test**, which is an exact probability test.*

The Necessity of Large Cell and Total Sample Sizes

Adhering to strict sample size requirements is paramount for ensuring the G-statistic reliably approximates the theoretical Chi-Square distribution. The generally accepted **rule-of-thumb** mandates two distinct conditions regarding sample size for optimal performance of the G-Test of Goodness of Fit.

Firstly, the count of observations (the observed frequency) in each individual category--or "cell"--must be reasonably large, with a minimum recommended threshold often set at 10 or more observations per cell. This threshold mitigates the risk of skewing the log-likelihood ratio calculation. Secondly, given the G-Test's preference for very large datasets, it is highly recommended to use this test primarily when the **total number of observations** in the study exceeds 1000. When these two conditions are met, the G-Test often provides a more accurate approximation to the Chi-Square distribution than the traditional Pearson's Chi-Square test, particularly for complex models or large tables.

ARABPSYCHOLOGY.COM

If these sample size constraints are violated, choosing a different statistical method is necessary to maintain validity. For instance, if you have insufficient cell counts but a dichotomous variable, an exact test is preferred. If you meet the cell count requirement but have a smaller total sample size, another asymptotic test might be more appropriate.

*If you have fewer than 10 observations in any cell, we recommend using the **Binomial Test** if your group variable has only two options, or the **Multinomial Goodness-of-Fit Test** if you have more than two categories. Furthermore, if you satisfy the minimum of 10 observations per cell but have fewer than 1000 total observations, we advise using the **One-Proportion Z-Test** for dichotomous variables or the **Chi-Square Goodness-of-Fit Test** if you have more than two categories.*

Illustrative Example of the G-Test Application

To illustrate the practical application of the G-Test of Goodness of Fit, consider a scenario involving demographic variables and known population distributions.

Variable of Interest: Gender (observed categories: **male/female**).

In this hypothetical scenario, the research question centers on whether our collected sample data reflects the national population gender split, which is hypothesized to be an equal 50-50 proportion (the expected frequencies). The formal statement guiding this investigation is the **null hypothesis**, which posits that there is absolutely no statistically significant difference between the proportion of females (or males) in our sample and the assumed 50% population proportion.

Before proceeding with the calculation, we confirm that our methodological assumptions have been met: the sample was collected via a random sample, the data points are independent, and the gender categories are mutually exclusive. Assuming the observed counts meet the minimum cell count threshold (e.g., >10 in both male and female groups) and the total sample size is sufficiently large, the calculation proceeds to yield a G-statistic and a corresponding p-value.

The resulting G-statistic measures the magnitude of the difference between observed and expected proportions. The associated p-value then quantifies the probability of observing our specific data (or data even more extreme) if the null hypothesis of a 50-50 split were truly correct in the population. The conventional threshold dictates that a p-value less than or equal to 0.05 indicates that the observed difference is sufficiently large to be considered statistically significant, allowing us to confidently reject the null hypothesis and conclude that our sample proportions deviate reliably from the expected 50-50 ratio.

ARABPSYCHOLOGY.COM