

How to Perform a G-Test to Analyze Categorical Data

Authored by
stats writer

January 22, 2026

RECOMMENDED CITATION

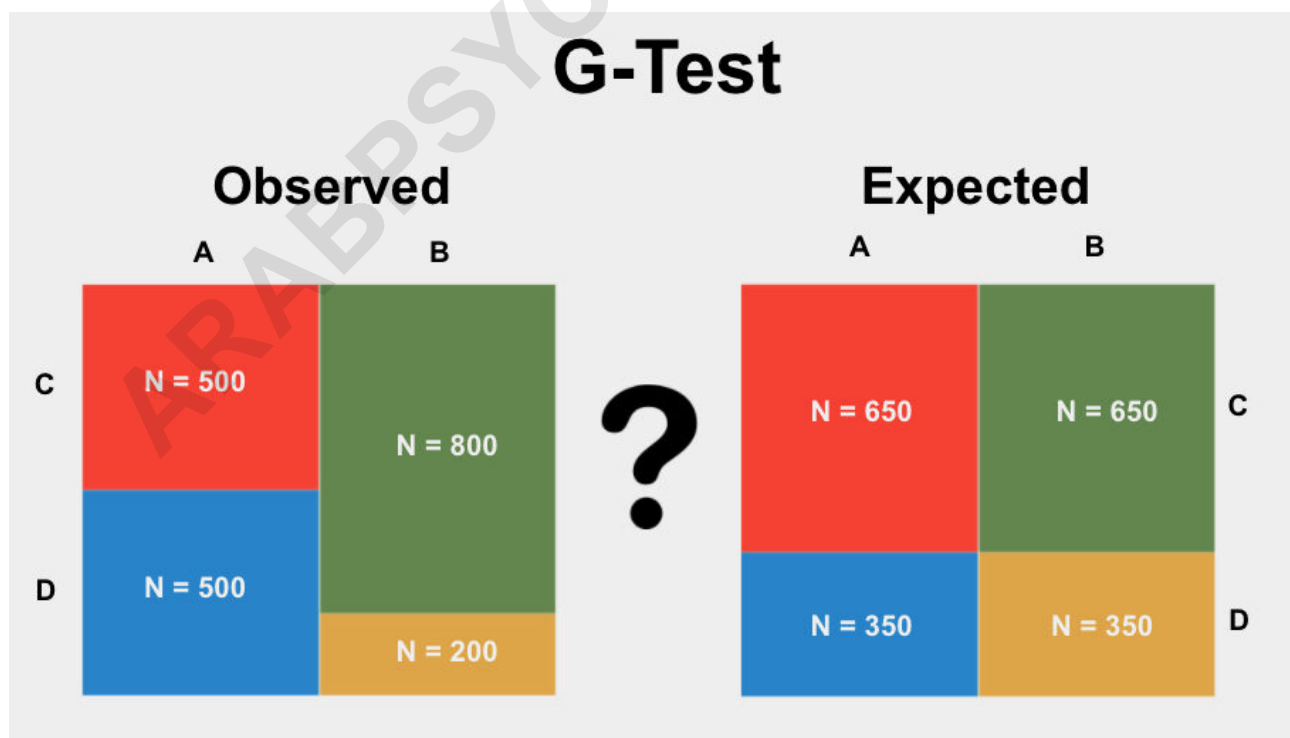
stats writer (2026). *How to Perform a G-Test to Analyze Categorical Data*.

PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=127082>

The G-Test, alternatively known as the G-Statistic or the G-Test of Independence, is a fundamental statistical test employed to assess whether there is a statistically significant disparity in distributions across two or more groups or categories. This powerful tool is widely utilized across diverse disciplines such as ecology, psychology, and data science for analyzing observational data. At its core, the test operates by comparing the observed frequencies of the data against the expected frequencies--those that would occur if the groups had no association. A larger G-Statistic value is indicative of a greater deviation between what was observed and what was expected, suggesting a meaningful relationship between the variables under scrutiny. Consequently, the G-Test provides crucial insights for research, hypothesis testing, and informed decision-making.

What is the G-Test?

The **G-Test** is a statistical test designed specifically to determine if the proportions or frequencies of categories across two or more categorical variables differ significantly from one another. It is often preferred in situations involving large datasets where the assumptions for other tests, like the Chi-Square test, might be less appropriate, particularly due to its relationship with the likelihood ratio principle. To effectively utilize the G-Test for independence, researchers generally require a minimum of two categorical variables, each possessing two or more distinct levels or options. Furthermore, for the approximation of the test statistic distribution to be reliable, a substantial total sample size, often exceeding 1000 observations, is recommended.



The G-Test is also widely recognized in statistical literature as the Likelihood Ratio Test, or

specifically, the *Log-Likelihood Ratio Test*.

Fundamental Assumptions for the G-Test

Like all rigorous statistical methods, the G-Test relies on several crucial assumptions regarding the nature and collection of the data. These assumptions are not merely theoretical formalities; they are prerequisite conditions that your dataset must satisfy for the resulting G-Statistic and associated p-value to be valid and accurately interpretable. Violating these core assumptions can lead to skewed results, inflated Type I or Type II errors, and ultimately, incorrect conclusions about the relationship between your variables.

Understanding and verifying these prerequisites is a vital step in the analytical process. The key assumptions necessary for the appropriate application of the G-Test include:

The data must originate from a **Random Sample**.

The observations must exhibit **Independence**.

The groups defined by the variables must be **Mutually Exclusive**.

Let us now explore the detailed implications of each of these three foundational requirements.

Random Sample

A cornerstone of inferential statistics is the requirement that the data points used in the analysis must be derived from a simple random sample of the population of interest. This means that every individual or unit within the population had an equal chance of being selected for the sample. If the sampling process is non-random or biased--for instance, if certain groups are systematically overrepresented or underrepresented--the resulting sample will not accurately reflect the true population parameters.

When the sample is not randomly determined, the analysis results are compromised. In statistical terminology, this situation introduces **selection bias**, which is a systematic tendency to produce inaccurate results due to flawed data collection methodology. Ensuring a true random sample is paramount to achieving external validity and allowing for reliable generalizations from the sample findings to the larger population.

Independence of Observations

The assumption of independence dictates that each individual observation or data point must be unrelated to all other observations within the dataset. In simpler terms, the value recorded for one unit of observation should not influence or be influenced by the value recorded for another unit. This is often a critical challenge in longitudinal studies or nested data structures.

This assumption is commonly violated in scenarios involving repeated measures over time from the same source (e.g., tracking a customer's behavior monthly, or measuring patient outcomes daily). Since data points originating from the same subject, customer, or experimental unit are inherently likely to be related or correlated, they violate the independence assumption. When observations are dependent, standard tests like the G-Test cannot accurately calculate the variance, leading to distorted test statistics and potentially false conclusions.

Mutually Exclusive Groups

For the G-Test to function correctly, the categories defining the categorical variables must be **mutually exclusive**. This means that any single unit of observation can belong to one and only one group within the variable structure. There must be no ambiguity or overlap in the definition of the groups.

For example, if you categorize survey respondents by their primary language (English, Spanish, French), a respondent can only select one. Similarly, if a variable tracks patient status as "Recovered" or "Non-Recovered," an individual cannot simultaneously occupy both categories. Maintaining mutually exclusive and exhaustive categories ensures that the observed frequencies are correctly counted and distributed within the contingency table structure upon which the G-Test is calculated.

Criteria for Applying the G-Test

Selecting the correct statistical test is paramount for sound research. The G-Test is specifically tailored for scenarios involving large datasets of categorical counts. Researchers should choose the G-Test when their analytical objective aligns with a specific set of five mandatory criteria, focusing on the nature of the variables, the goal of the analysis, and the sample size characteristics.

You should primarily use the G-Test when the following conditions are met simultaneously:

The primary goal is to test for a significant **Difference** or association between two variables.

The variables of interest are inherently **Proportional or Categorical**.

The variables contain **Two or More Options** (levels) for classification.

The data consists of **Independent Samples** or observations.

The overall sample size is substantial, ideally having **More than 1000 values in total**.

A deeper understanding of these requirements will clarify the appropriate context for employing the G-Test.

Testing for Difference or Association

The G-Test is specifically classified as a test of association or independence. Its purpose is to determine whether the distribution of one categorical variable changes significantly across the levels of another categorical variable. Essentially, you are seeking a statistical answer to the question: Does the outcome differ significantly between the groups?

This contrasts sharply with other statistical objectives, such as testing for a linear or curvilinear relationship (correlation), or building a model to predict the value of one variable based on others (regression or prediction analysis). If your primary hypothesis involves determining if proportions are distinct across groups, the G-Test is a strong candidate.

Requirement for Categorical Data

The fundamental requirement for the G-Test is that the variables under investigation must be nominal, ordinal, or counts derived from such scales--meaning they must be categorical variables. A categorical variable classifies observations into distinct, non-overlapping categories without a natural numerical order (e.g., eye color, region, or product type).

Furthermore, the test applies to data represented as proportional variables. Proportions are derived directly from counts within categorical variables, such as calculating conversion rates (10% success vs. 15% success), the percentage of subjects who voted versus those who abstained, or the fraction of agricultural yield that survived a specific treatment. If your analysis involves continuous numerical variables (e.g., height, temperature, income), different statistical procedures are required.

If the variables you wish to compare are continuous in nature, you should consider using an Independent Samples T-Test or analysis of variance (ANOVA) instead.

Defining the Number of Categories

The G-Test is highly versatile in accommodating various study designs because it does not limit the number of categories within the variables. Your categorical variables must simply possess at least two possible options, or levels. Examples of such variables include binary outcomes (e.g., made a purchase: yes/no; disease recovery: present/absent) or multinomial outcomes (e.g., preferred color: red/blue/green; satisfaction level: low/medium/high).

The Requirement for Independent Samples

Crucially, the observations being compared must come from **independent samples**. This concept reinforces the assumption of independence mentioned earlier but is often discussed separately to

stress the study design. Independent samples imply that the data collected in one group has absolutely no systemic relationship or pairing with the data collected in the other group.

A classic violation occurs with paired or matched designs, where the same individuals or units are measured under two different conditions or at two points in time. Since measurements from the same person are inherently correlated, they are not independent samples. The G-Test is designed for situations where the groups are distinct and unrelated (e.g., comparing Treatment Group A patients vs. Treatment Group B patients).

If your design involves repeated measures or paired observations from a single sample, you should instead consider employing tests specifically designed for dependent samples, such as the McNemar Test.

Minimum Sample Size Considerations

While there is debate regarding the exact minimum cell counts for the G-Test versus the Chi-Square test, the G-Test tends to perform better than the traditional Chi-Square test, especially when dealing with very large datasets. As a practical rule of thumb for robust results, particularly in complex contingency tables, it is highly recommended to use this statistical test when the total number of observations (N) is 1000 or greater.

This substantial sample size helps ensure that the theoretical assumptions underlying the chi-squared approximation of the G-Statistic distribution are met, leading to more accurate probability estimates (p-values). When sample sizes are small or cell counts are low, alternative exact tests are necessary to maintain statistical validity.

For instances where you have very small cell counts (fewer than 10 in any cell), we strongly recommend using Fisher's Exact Test. Conversely, if you have moderate sample sizes (fewer than 1000 total observations) but adequate cell counts (more than 10 in every cell), and only two options in your categorical variable, the Two-Proportion Z-Test is often preferred. If you have moderate total observations and more than two options in your categorical variable, the Chi-Square Test of Independence is the standard recommendation.

Illustrative G-Test Example

To concretely illustrate the application of the G-Test, consider a study focused on comparing the effectiveness of two novel medical treatments:

Primary Grouping Variable: Treatment Type (Levels: A or B)

Outcome Variable (Categorical): Recovered from disease (Levels: Yes or No)

In this scenario, the researchers are fundamentally interested in investigating whether the two distinct treatment groups exhibit significantly different rates of recovery from the disease. The formal statistical starting point is the **null hypothesis**, which posits that there is absolutely no difference in recovery proportions between Treatment A and Treatment B; any observed differences are merely due to random sampling variation.

Since the outcome variable is binary with only two possible values (Yes/No), and all other prerequisites regarding independence, randomization, and sample size are met, the G-Test is the appropriate analytical instrument. It is important to remember that the G-Test's utility extends beyond binary variables; it would be equally suitable if the outcome variable had multiple levels, such as "Full Recovery," "Partial Recovery," and "No Recovery."

The result of the analysis will yield the G-Statistic itself, alongside a critical probability value known as the p-value. This p-value quantifies the probability of observing data as extreme as (or more extreme than) the collected sample data, assuming the null hypothesis (no difference in recovery rates) is actually true. If the calculated p-value is less than or equal to the predetermined alpha level (typically 0.05), the result is deemed **statistically significant**. This significance allows the researchers to confidently reject the null hypothesis and conclude that the difference in recovery rates observed between the two treatment types is genuine and highly unlikely to be attributable to chance alone.