How to Calculate G-Test of Goodness of Fit?

Authored by

stats writer

December 6, 2025

Introduction to the G-Test of Goodness of Fit

The G-Test of Goodness of Fit is a powerful and increasingly popular statistical methodology employed to determine if the frequency distribution of a sample significantly deviates from a theoretically expected distribution. Fundamentally, this test operates on categorical data, where observations fall into discrete categories, and the core objective is to compare the actual counts (the observed frequencies) against the counts hypothesized under a null model (the expected frequencies). Unlike traditional tests that rely on calculating squared differences, the G-test is rooted in the principles of likelihood ratio tests, providing specific statistical properties that make it advantageous in complex modeling scenarios, particularly in ecological and biological research.

The primary appeal of the G-test lies in its derivation from the maximum likelihood theory, making it asymptotically equivalent to the Chi-Squared Test (chi^2) as sample sizes increase. However, the G-test often provides a better approximation to the theoretical chi-squared distribution, especially when dealing with smaller expected counts or when performing post-hoc analysis where data is being partitioned or combined. Understanding this test is crucial for researchers who need to assess whether a theoretical model, such as Mendelian genetics ratios, a uniform distribution, or a known population prevalence, adequately explains the distribution observed in their experimental data.

When applying the G-test, the resulting statistic (G) quantifies the magnitude of the discrepancy between what was observed and what was anticipated. A small G value suggests that the observed data aligns closely with the null hypothesis expectations, implying a good fit. Conversely, a large G value signifies a significant departure, indicating that the observed distribution is unlikely to have arisen by chance under the null model. This statistical measure, therefore, serves as a vital tool for validating or rejecting theoretical hypotheses concerning data distribution.

The Theoretical Basis: Likelihood Ratio Testing

The G-test is formally known as a log-likelihood ratio test, distinguishing it sharply from the sum-of-squares approach utilized by the Chi-Squared Test. Its calculation is fundamentally based on comparing the likelihood of the observed data arising under the null hypothesis (H_0) versus the likelihood of the data arising under the saturated model (the model that perfectly fits the observed data). This comparison is performed using the ratio of these two likelihoods, which is then transformed using the natural logarithm to yield the G statistic. This derivation provides the G-test with superior additive properties, meaning that G statistics calculated for independent studies or different components of a contingency table can be summed together to yield an overall G statistic for the combined data.

In mathematical terms, the G statistic is proportional to the difference between the likelihood of the observed frequencies and the expected frequencies, where this difference is scaled by the natural

log of the ratio of these frequencies. This log transformation is the key mechanism that allows the test to leverage the powerful mathematical framework of likelihood theory. The underlying concept is that if the null hypothesis is true, the ratio of the likelihoods will be close to one, making the natural logarithm close to zero, resulting in a small G statistic. If the null hypothesis is false, the ratio deviates significantly from one, leading to a much larger G value.

The formal connection between the G-test and the likelihood principle is what endows it with its statistical rigor. Because the G-test statistic is asymptotically distributed as a chi-squared random variable, we can use the standard chi-squared distribution tables and associated degrees of freedom to determine the corresponding P-value. This P-value then dictates the decision regarding the null hypothesis, providing a quantified measure of evidence against the expected model. This robust theoretical grounding makes the G-test a preferred choice in fields like information theory and model selection, where likelihood comparisons are central to the analysis.

Formula and Step-by-Step Calculation

Calculating the G-test statistic requires meticulous attention to the observed and expected counts within each category. The general formula for the G-test for goodness of fit is:

G = 2 * Σ

Where $Sigma$ represents the summation across all categories i; $Observed_i$ is the actual count in category i; $Expected_i$ is the hypothesized count in category i; and ln represents the natural logarithm. The factor of two ensures that the resulting statistic adheres to the chi-squared distribution under the null hypothesis. The systematic calculation process is critical for obtaining an accurate result and drawing valid statistical conclusions.

The calculation process can be broken down into specific steps, ensuring clarity and correctness.

Establish Hypotheses: Define the Null Hypothesis (H_0), stating that the observed distribution is the same as the expected distribution, and the Alternative Hypothesis (H_A), stating that the distributions differ.

Determine Expected Frequencies: Calculate the expected count (E_i) for each category based on the total sample size (N) and the hypothesized probabilities (p_i) for that category (i.e., $E_i = N times p_i$).

Calculate the Ratio and Logarithm: For each category, calculate the ratio of the observed to the expected frequencies (O_i / E_i) and then take the natural logarithm of this ratio: $ln(O_i / E_i)$.

Calculate the Contribution of Each Category: Multiply the observed frequency by the calculated

logarithm from the previous step: $O_i times ln(O_i / E_i)$.

Sum and Finalize: Sum the results from Step 4 across all categories, and then multiply the total sum by two. This yields the final G statistic.

Determine Degrees of Freedom: The degrees of freedom (df) are calculated as the number of categories minus the number of parameters estimated from the data. In the simplest goodness-of-fit case, df is typically $k - 1$, where k is the number of categories.

It is essential to note that the G-test, like the Chi-Squared Test, performs best when expected counts are not excessively small. While conventional guidelines often suggest minimum expected counts (e.g., all expected counts should be greater than 5), the G-test tends to be slightly more robust than chi^2 when these conditions are marginally violated, though caution must always be exercised with very sparse data.

Assumptions and Conditions for Application

For the results of the G-Test of Goodness of Fit to be statistically valid, several fundamental assumptions regarding the data structure and collection process must be met. Ignoring these prerequisites can lead to spurious conclusions and misinterpretation of the significance level. The assumptions mirror those required for any asymptotic test relying on the chi-squared distribution approximation.

First, the data must consist of **count data**--that is, the frequencies or totals of independent events observed in each category. The G-test is inappropriate for continuous data or interval data unless it has been categorized into discrete bins. Second, the observations must be **independent**. This means the outcome of one observation or event should not influence the outcome of any other observation. If the sampling method involves dependencies (e.g., repeated measures on the same individual), the G-test is not the correct analytical tool, and alternative methods such as generalized linear mixed models (GLMMs) might be necessary.

The third critical condition relates to the **sample size and expected frequencies**. The test relies on the asymptotic property that the G statistic follows a chi-squared distribution. This approximation holds true only when the sample size is sufficiently large. Specifically, it is generally required that no more than 20% of the expected frequencies (E_i) are less than 5, and none of the expected frequencies should be zero. If these conditions are severely violated, the P-value derived from the chi-squared distribution will be inaccurate, potentially leading to Type I or Type II errors. In such cases, methods like Fisher's exact test (for small tables) or combining categories (if logically justifiable) should be considered.

Comparison with the Chi-Squared Test

The G-test and the Chi-Squared Test are often used interchangeably for goodness-of-fit applications, as both tests converge to the same result for large sample sizes, and both use the same chi-squared distribution for hypothesis testing. However, the G-test possesses distinct advantages that make it preferred in certain statistical contexts, particularly when dealing with complex experimental designs or when maximum likelihood estimation is paramount.

One of the primary differences lies in their mathematical foundation. The Chi-Squared statistic measures the difference between observed and expected counts using squared Euclidean distance: $chi^2 = Sigma $. In contrast, the G-test utilizes the natural logarithm, derived from the theory of the Likelihood Ratio Test. This logarithmic structure provides the G-test with superior mathematical properties, most notably its additivity. When dealing with hierarchical models or complex experimental designs, the G-test allows researchers to partition the total variability (total G statistic) into components attributable to different sources of variation. This partitioning ability is highly valuable in analysis of variance (ANOVA) type applications for categorical data.

Furthermore, in situations where statistical power is marginal or the expected cell counts are moderate, the G-test generally provides a closer approximation to the true chi-squared distribution than the chi^2 test. For these reasons, the G-test is sometimes recommended by statisticians as the default choice over the traditional Pearson's Chi-Squared Test, especially in contemporary statistical software packages. However, due to the historical prevalence and computational simplicity of the Chi-Squared Test, it remains the standard introductory test for goodness of fit. Researchers should select the G-test when analyzing structured data or when the additive property is required for detailed decomposition of variance.

Interpreting the G-Statistic and P-Value

The interpretation of the G-statistic follows the same logical framework as interpreting the Chi-Squared statistic. The G value itself is not directly interpretable in terms of the effect size, but rather serves as a measure of the likelihood ratio. A large G value signifies a poor fit between the observed frequencies and the hypothesized expected frequencies, suggesting that the null hypothesis (H_0) of no difference is implausible.

To convert the G statistic into a probabilistic statement, it must be compared against the critical values of the chi-squared distribution using the appropriate degrees of freedom (df). The P-value derived from this comparison represents the probability of observing a G statistic as extreme or more extreme than the one calculated, assuming that the null hypothesis is true. If the P-value is less than the predetermined significance level (alpha, typically 0.05), the researcher rejects the null hypothesis, concluding that there is a statistically significant lack of fit between the observed data and the theoretical model.

Conversely, if the P-value is greater than the chosen alpha level, the researcher fails to reject the null hypothesis. This outcome suggests that the observed deviations are small enough to be attributed to random chance, and the theoretical model provides an adequate explanation for the observed distribution of the categorical data. It is crucial to remember that failing to reject H_0 does not prove the model is correct, but merely suggests that the data does not provide sufficient evidence to conclude it is wrong. Furthermore, researchers should always inspect the residuals (the differences between O_i and E_i) to understand exactly which categories contribute most significantly to the overall lack of fit.

Practical Implementation and Software Usage

While the calculation of the G-statistic is straightforward, performing the test manually for large datasets can be cumbersome and prone to computational errors. Consequently, statistical software packages are routinely employed for the G-Test of Goodness of Fit. Packages like R, SAS, SPSS, and Python libraries (e.g., SciPy) offer functions that automate the calculation of the G statistic, the degrees of freedom, and the corresponding P-value.

For instance, in R, the G-test is often incorporated within the standard statistical testing functions, frequently as an option within the contingency table analysis tools or sometimes explicitly through specialized packages. When using these functions, the user typically provides a vector of observed frequencies and a vector of expected probabilities or expected frequencies. The software then handles the logarithmic calculation and the subsequent comparison to the chi-squared distribution.

Using software ensures not only accuracy but also efficiency, allowing researchers to rapidly assess multiple hypotheses or analyze large datasets. When reporting results, it is standard practice to state the G statistic, the degrees of freedom, and the P-value, often accompanied by the total sample size and a visualization (such as a bar chart) comparing the observed and expected counts to aid interpretation. Proper utilization of these tools ensures that the statistical inference drawn is both reliable and reproducible, adhering to the best practices in quantitative research.

Advantages and Limitations of the G-Test

The G-test offers several marked advantages over its statistical cousins, primarily stemming from its derivation based on the maximum likelihood principle. The core advantage is the **additivity of the G statistics**, which is unmatched by the Pearson chi^2 statistic. This makes the G-test essential for complex experimental designs, such as those involving log-linear models, where researchers need to decompose interaction effects or test nested hypotheses efficiently. Furthermore, its connection to the Likelihood Ratio Test makes it theoretically superior for modeling applications where likelihood maximization is the goal.

Another significant benefit is its generally improved performance for smaller samples or slightly lower expected frequencies compared to the standard Chi-Squared Test. While both tests require large sample approximations, the G-test tends to approach the theoretical chi^2 distribution more rapidly, resulting in more accurate P-values under marginal conditions. This statistical refinement makes the G-test a robust choice in disciplines like population genetics or ecology where sample collection might be constrained.

However, the G-test is not without limitations. Like the Chi-Squared Test, it is sensitive to low expected frequencies. If the expected count in any category is very small (approaching zero), the term $ln(O_i / E_i)$ becomes unstable, potentially leading to inflated G statistics and unreliable P-values. In these severe cases of sparse data, techniques such as collapsing categories or employing Monte Carlo simulation methods are necessary. Additionally, because the G-test is less commonly taught in introductory statistics compared to chi^2, some practitioners may be less familiar with its application and interpretation, leading to potential misapplication if the underlying likelihood theory is not fully grasped.