MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Review Article

Open Access

CrossMark

Using the one-sample chi-square (goodness-of-fit) test correctly in health research: frequent mistakes and practical guidance

Abstract

Categorical data are central to clinical and epidemiological research, and appropriate statistical analysis is essential for drawing defensible conclusions. The one-sample chi-square test, also called the goodness-of-fit test, is used to assess whether observed frequencies are compatible with a specified expected distribution, such as equal group proportions, established genetic ratios, or a benchmark derived from previous research. Despite its broad use, the test is sometimes applied to inappropriate data or without adequate attention to its assumptions. This paper reviews common errors in health research, including applying the test to continuous measurements, overlooking small expected counts, violating independence, confusing goodness-of-fit with tests of association, and interpreting p-values without considering effect size or substantive importance. A hypothetical clinical example illustrates dataset preparation, step-by-step execution in SPSS, reporting of results, and cautious interpretation. Throughout the paper, we emphasise that a p-value describes the compatibility of the observed data with the null model, conditional on the assumptions of the statistical procedure; it does not represent the probability that the null hypothesis is true or that an observed result is due to chance.

Volume 15 Issue 1 - 2026

Ilker Etikan
Faculty of Medicine, Head of the Biostatistics Department, Near East Avenue, Turkey

Correspondence: Ilker Etikan, Faculty of Medicine, Head of the Biostatistics Department, Near East Avenue, Postcode 99138, Nicosia, North Cyprus, Mersin 10, Turkey

Received: September 15, 2026 | **Published:** September 29, 2026

Introduction

A large share of clinical and biomedical research involves categorical variables, including diagnostic categories, tumour grade, treatment response, and genetic variant type. When the research question concerns whether the distribution of a single categorical variable is compatible with a specified reference distribution—for example, a historical rate, a Mendelian inheritance ratio, or an assumption of equal proportions—the one-sample chi-square goodness-of-fit test is a conventional analytical tool.[1,2]

The statistical theory underlying the test is well established, but application errors remain common. These errors include treating continuous measurements as frequency counts, applying the test when expected counts are sparse, treating repeated observations as independent, and using a goodness-of-fit test to assess the association between two variables. A further interpretive problem occurs when statistical significance is treated as evidence of clinical importance or when a p-value is described as the probability that a finding is due to chance. This paper provides practical guidance intended to help health researchers recognise these problems and report one-sample chi-square analyses more precisely.

Common errors and recommended approaches

Feeding continuous measurements into the test

The one-sample chi-square goodness-of-fit test is designed for frequency counts in discrete categories. Continuous measurements such as age, blood pressure, hemoglobin concentration, or tumor size are therefore not appropriate in their original continuous form. The relevant unit of analysis is the number of observations falling within each category, not the individual measurement itself.

Recommended approach: Continuous variables should be analysed using statistical methods appropriate to their distribution and research question, such as the t-test or ANOVA, or suitable non-parametric alternatives. If a chi-square analysis is scientifically justified, a continuous variable may first be categorised into clinically meaningful groups—for example, hypokalemia (<3.5 mEq/L), normal range (3.5–5.0 mEq/L), and hyperkalemia (>5.0 mEq/L). Categorization should be justified because it can discard information.

Ignoring the expected-count requirement

The chi-square approximation is most reliable when expected counts are sufficiently large. Small samples, rare conditions, or classifications divided into many categories can produce sparse expected frequencies. A commonly used rule for a one-sample chi-square test is that expected counts should be at least 5 in each category.[1,2]

Recommended approach: Before interpreting the test, calculate the expected counts. Where scientifically appropriate, sparse categories may be combined into broader, clinically sensible groups, or a larger sample may be required. Category collapsing should not be performed solely to obtain a significant or non-significant p-value.

Overlooking the independence requirement

The standard one-sample chi-square procedure assumes that observations are independent. This assumption is violated when the same participant contributes multiple observations that are analysed as if they were separate independent cases. For example, if 50 hypertensive patients are assessed at baseline, one month, and three months, the repeated measurements cannot simply be entered into a standard one-sample chi-square analysis as independent observations.

Recommended approach: Repeated or paired measurements require methods designed for dependent data. For binary outcomes measured at two time points, McNemar's test may be appropriate; for binary outcomes measured at three or more time points, Cochran's Q test may be considered.[3] The appropriate method ultimately depends on the study design and outcome structure.

Confusing goodness-of-fit with a test of association

A one-sample goodness-of-fit test and a test of association answer different questions. The goodness-of-fit test assesses whether the observed proportions of a categorical variable are consistent with a specified reference distribution. It does not test whether two categorical variables are related.

For example, asking whether smoking status is associated with COPD involves two categorical variables and is not a one-sample goodness-of-fit question.

Recommended approach: Questions about relationships between two categorical variables generally require a test of independence or another method appropriate to the study design. Epidemiological measures such as odds ratios or relative risks may also be relevant when the research question concerns association or risk (McHugh, 2013).[4]

Overreliance on the P-value in large datasets

In large datasets, relatively small deviations between observed and expected proportions can produce small p-values because the chi-square statistic is sensitive to sample size. A statistically significant result, therefore, does not by itself establish that the observed deviation is clinically or practically important.

Recommended approach: Report the magnitude of the deviation alongside the p-value. A standardised effect-size measure, such as Cramér's V, can provide additional information about the size of the discrepancy. The interpretation should also consider the clinical context, study design, measurement quality, and precision of the estimates.[5,6]

Quick decision checklist

Before using a one-sample chi-square goodness-of-fit test, researchers should consider the following questions:

Question	Expected answer
Is the variable genuinely categorical?	Yes
Does each participant contribute one independent observation?	Yes
Are the expected counts sufficiently large (commonly ≥5)?	Yes
Is the aim to compare observed proportions with a specified reference distribution?	Yes

When these conditions are satisfied, a one-sample chi-square goodness-of-fit analysis may be appropriate. If one or more conditions are not satisfied, the research question and study design should be reconsidered before selecting the test.

Worked clinical example

Scenario and purpose

The following example is hypothetical and is provided solely to illustrate the mechanics and interpretation of a one-sample chi-square goodness-of-fit test. It should not be interpreted as evidence from an actual phase III clinical trial.

Suppose that historical data for established first-line antihypertensive agents indicate an illustrative side-effect distribution among symptomatic patients of approximately 50% headache, 30% dizziness, and 20% nausea. In a hypothetical dataset of 200 patients who experienced a side effect while receiving a new antihypertensive agent, the question is whether the observed distribution is compatible with this specified historical distribution.

Side-effect category	Observed (O)	Expected proportion	Expected (E)
Headache	85	0.5	100 (200 × 0.50)
Dizziness	75	0.3	60 (200 × 0.30)
Nausea	40	0.2	40 (200 × 0.20)
Total	200	1	200

The expected counts are all at least 5, the outcome is categorical, and each hypothetical patient contributes one observation to one category. These conditions make the example suitable for demonstrating the one-sample goodness-of-fit procedure.

Analysis procedure in SPSS

Defining the data and weighting cases

Using summarized frequencies, the analysis can be set up in IBM SPSS as follows.[7]

In Variable View, create two variables: Side_Effect (numeric, nominal) and Patient_Count (numeric, scale).

For Side_Effect, assign codes 1 = Headache, 2 = Dizziness, and 3 = Nausea.

In Data View, enter the summary counts: Side_Effect = 1 with Patient_Count = 85; Side_Effect = 2 with Patient_Count = 75; Side_Effect = 3 with Patient_Count = 40.

Select Data > Weight Cases, choose Weight cases by, move Patient_Count into the Frequency Variable box, and click OK.

SPSS will then use the frequency variable to represent the summarized observations.

Executing the goodness-of-fit test

Go to Analyse> Nonparametric Tests > Legacy Dialogues> Chi-Square.

Move Side_Effect into the Test Variable List.

Turn off the default option for equal expected proportions.

Select Values and enter the expected proportions in category order: 0.50, 0.30, and 0.20.

Open Options if descriptive statistics are required, then click Continue and OK.

Results and reporting

Main results

For the illustrative dataset, the one-sample chi-square goodness-of-fit test gives $\chi^2(2) = 6.00$, $p = .0498$. With a pre-specified significance level of $\alpha = .05$, this p-value is below the conventional threshold. More precisely, under the null model and the assumptions of the test, a result at least as discrepant as the observed result has probability 0.0498 according to the chi-square reference distribution.

The observed frequencies were 85 headache cases versus 100 expected, 75 dizziness cases versus 60 expected, and 40 nausea cases versus 40 expected. Thus, the largest deviations in this illustrative dataset were a lower-than-expected frequency of headache and a higher-than-expected frequency of dizziness.

Citation: Etikan I. Using the one-sample chi-square (goodness-of-fit) test correctly in health research: frequent mistakes and practical guidance. *Biom Biostat Int J.* 2026;15(1):69–72. DOI: 10.15406/bbij.2026.15.00450

Side-effect category	Observed (O)	Observed %	Expected proportion	Expected (E)	Residual (O − E)
Headache	85	42.50%	0.5	100	−15.0
Dizziness	75	37.50%	0.3	60	15
Nausea	40	20.00%	0.2	40	0
Total	200	100.00%	1	200	—

Note: N = 200. All expected counts are ≥5.

Sample write-up

To assess whether the illustrative side-effect distribution was compatible with the specified historical distribution (50% headache, 30% dizziness, and 20% nausea), a one-sample chi-square goodness-of-fit test was conducted. The observed distribution differed from the specified distribution, $\chi^2(2) = 6.00$, $p = .0498$. Headache occurred less often than expected (42.5% vs 50.0%), whereas dizziness occurred more often than expected (37.5% vs 30.0%); nausea was equal to the expected proportion (20.0%). Because this is a hypothetical example, these results provide only statistical interpretation and do not provide evidence regarding the safety or clinical effectiveness of an actual antihypertensive agent.

Interpretation

The statistical result should be interpreted in relation to the null model rather than as a probability that the null hypothesis is true or that the observed difference was caused by chance. In this example, the p-value of .0498 indicates that the observed discrepancy between the sample distribution and the specified historical distribution is relatively incompatible with the null model, conditional on the assumptions of the test.

The p-value does not establish the magnitude, clinical importance, causation, or safety implications of the observed differences. Those questions require additional evidence, appropriate study designs, and, where relevant, estimates of effect and uncertainty.

Effect Size and Practical Interpretation

For large clinical datasets, an effect-size measure can complement the p-value. In this illustrative example, the reported Cramér's V is calculated as follows:

$$V = \sqrt{[\chi^2 / (N \times (k - 1))]}$$

where χ^2 is the chi-square statistic, N is the total number of independent observations, and k is the number of categories.

Using $\chi^2 = 6.00$, $N = 200$, and $k = 3$:

$$V = \sqrt{[6.00 / (200 \times 2)]} = \sqrt{0.015} \approx 0.122.$$

This value indicates that the standardised discrepancy is not large. However, effect-size interpretation should not rely on an isolated numerical cut-off; the substantive meaning of the deviation depends on the research context and the consequences of the outcome. In particular, statistical significance and practical or clinical importance are distinct concepts.

Formal hypotheses

For the illustrative example, the hypotheses can be stated as:

Null hypothesis (H_0): P(headache) = 0.50, P(dizziness) = 0.30, and P(nausea) = 0.20.

Alternative hypothesis (H_1): At least one category probability differs from its specified historical value.

The alternative hypothesis does not identify which category differs in the population; the observed residuals can be used descriptively to show where the sample distribution departs from the specified proportions.

Discussion

Main findings and statistical interpretation

The worked example demonstrates how a one-sample chi-square goodness-of-fit test can be used to compare an observed categorical distribution with a specified reference distribution. The illustrative result, $\chi^2(2) = 6.00$, $p = .0498$, indicates a discrepancy from the specified distribution under the test's assumptions. The observed proportions show fewer headache cases and more dizziness cases than expected, while the nausea proportion matches the reference value.

These findings should not be interpreted as evidence that the hypothetical drug causes more dizziness, increases the risk of falls or fractures, or should replace an established treatment. The analysis contains no comparative clinical trial design, adjustment for patient characteristics, dose information, confidence intervals for clinical effects, or independent evidence about downstream outcomes such as falls or fractures. Such clinical conclusions would require appropriate empirical evidence beyond this goodness-of-fit calculation.

Clinical considerations: scope of the example

In a real clinical study, differences in side-effect distributions could be considered alongside patient age, comorbidities, concomitant medications, dose, exposure time, and other clinically relevant factors. For example, dizziness could be clinically relevant for patients with a history of falls, but this example does not measure falls or establish a causal relationship between dizziness and falls. Accordingly, these issues are best presented as considerations that could motivate further investigation rather than as conclusions supported by the present illustrative analysis.

Limitations and next steps

The example has several limitations by design. First, it uses a hypothetical dataset and therefore cannot establish clinical effectiveness or safety. Second, it considers the distribution of one reported side-effect category at a time and does not model patient-level factors that may influence adverse events. Third, assigning each patient to a single primary symptom would not capture participants who experienced multiple symptoms simultaneously. In an actual clinical research programme, larger and appropriately designed studies could use contingency-table methods, regression models, or other methods suited to the specific outcome and study design.

Citation: Etikan I. Using the one-sample chi-square (goodness-of-fit) test correctly in health research: frequent mistakes and practical guidance. *Biom Biostat Int J.* 2026;15(1):69–72. DOI: 10.15406/bbij.2026.15.00450

Practical Recommendations

Use the one-sample chi-square goodness-of-fit test only when the outcome is categorical, and the research question concerns comparison with a specified reference distribution.

Check expected counts before interpreting the chi-square approximation and address sparse categories using a scientifically justified strategy.

Ensure that each observation represents an independent unit for the standard test; use methods for dependent data when observations are repeated or paired.

Do not use a one-sample goodness-of-fit test to answer questions about associations between two categorical variables.

Report the chi-square statistic, degrees of freedom, exact or appropriately rounded p-value, observed and expected frequencies, and a relevant effect-size measure when useful.

Interpret p-values as measures of compatibility between the observed data and the null model, conditional on the assumptions of the analysis—not as probabilities that the null hypothesis is true or that the result is due to chance.

Separate statistical findings from clinical conclusions. In applied health research, clinical implications should be supported by the study design and relevant evidence rather than inferred solely from statistical significance.

Appendix: calculation of Cramér's V for the example

For the illustrative dataset:

$\chi^2 = 6.00$, $N = 200$, $k = 3$

$V = \sqrt{[6.00 / (200 \times (3 - 1))]} \approx 0.122$.

The value is reported as a standardised description of the discrepancy in this example. It should be interpreted together with the observed proportions, the research context, and the design, rather than used as a stand-alone basis for clinical judgment.

Acknowledgements

None.

Conflicts of interest

The author declares that there is no conflicts of interest.

References

1. Agresti A. *An introduction to categorical data analysis*. 3rd edition. John Wiley & Sons; 2018.
2. Campbell MJ, Swinscow TDV. *Statistics at square one*. 11th edition. Wiley-Blackwell; 2011.
3. Cochran WG. Some methods for strengthening the common χ^2 tests. *Biometrics*. 1954;10(4):417–451.
4. McHugh ML. The chi-square test of independence. *Biochem Med (Zagreb)*. 2013;23(2):143–149.
5. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd edition. Lawrence Erlbaum Associates; 1988.
6. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337–350.
7. Field A. *Discovering statistics using IBM SPSS statistics*. 5th edition. SAGE Publications; 2018.

Citation: Etikan I. Using the one-sample chi-square (goodness-of-fit) test correctly in health research: frequent mistakes and practical guidance. *Biom Biostat Int J*. 2026;15(1):69–72. DOI: 10.15406/bbij.2026.15.00450