

How to Perform a McNemar Test for Comparing Paired Data

Authored by
stats writer

January 22, 2026

RECOMMENDED CITATION

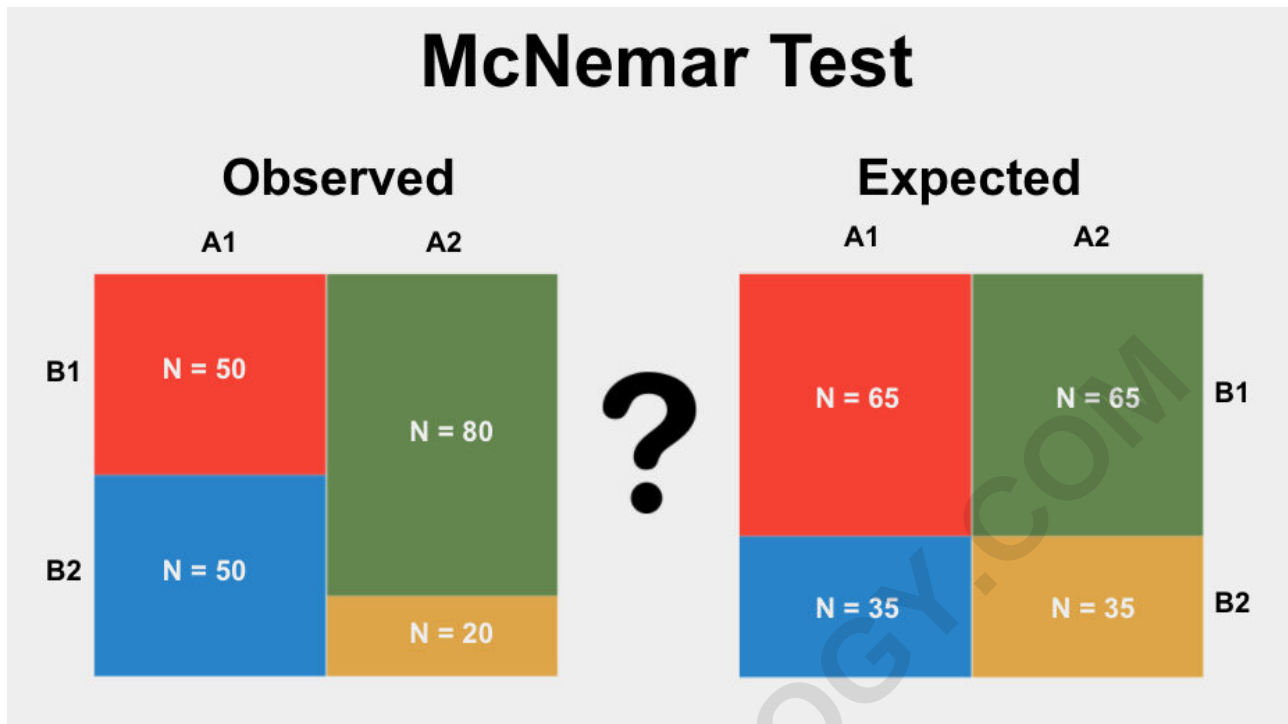
stats writer (2026). *How to Perform a McNemar Test for Comparing Paired Data*.
PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=127058>

The McNemar Test stands as a powerful, non-parametric statistical tool specifically designed to analyze the significance of changes in proportions between two related groups or measurements. It is fundamental in scenarios where the subjects act as their own controls, such as before-and-after studies, cross-over trials, or matched-pair case-control designs. Unlike tests for independent samples, the McNemar Test focuses strictly on **discordant pairs**--those subjects whose outcomes changed between the two observations--to assess whether an intervention or passage of time has caused a statistically significant shift in outcomes.

This versatile test is commonly employed across diverse research fields, including **medicine**, **psychology**, and the **social sciences**, to rigorously compare the effectiveness of different treatments, assess diagnostic accuracy, or analyze shifts in opinion or behavior. It is particularly valuable because it leverages the underlying chi-square distribution, making it robust even when dealing with relatively smaller sample sizes, provided the data structure is appropriately paired. By focusing on the magnitude and direction of change within matched pairs, the McNemar Test provides researchers with a robust methodology to determine if observed differences are genuinely statistically significant or merely attributable to random chance.

Defining the McNemar Test and Its Purpose

The McNemar Test is formally defined as a non-parametric statistical test used to evaluate whether the marginal frequencies (or proportions) of categories across two related groups differ significantly from each other. This methodology is crucial when analyzing dichotomous (binary) data collected from the same individuals or from matched pairs. The structure of the data typically involves organizing the results into a 2x2 contingency table, highlighting the joint outcomes of the two related measurements (e.g., success before vs. success after).



The McNemar Test is also known by several descriptive alternative titles, including the **Paired Sample Z-Test**, the **Paired Sample Proportion Z-Test**, and simply the **Paired Proportion Z-Test**. These names emphasize the test's application to paired or dependent data structures where the goal is to compare proportions.

The primary goal of the McNemar Test is to test the null hypothesis that the marginal proportions are equal, meaning that the probability of shifting from outcome A to outcome B is equal to the probability of shifting from outcome B to outcome A. If the test rejects this null hypothesis, researchers can conclude that there is a significant effect or change between the two time points or conditions. It is essential to recognize that this test specifically ignores the pairs that show consistent results (e.g., Yes/Yes or No/No), focusing entirely on the discordant outcomes (Yes/No or No/Yes) as they represent the true change. The calculation focuses solely on these off-diagonal cells to measure the asymmetry of change.

When performing the calculation, the McNemar statistic approximates the chi-square distribution with one degree of freedom. This method is particularly sensitive to differences in proportions, allowing researchers to efficiently analyze matched-pair data without resorting to computationally intensive permutation tests, provided the sample size is adequately large, especially within the discordant cells. For scenarios involving smaller sample sizes, the exact binomial test equivalent of the McNemar test is often preferred to maintain accuracy.

Critical Assumptions for the Valid Application of the McNemar Test

Like all inferential statistical procedures, the McNemar Test relies on a set of critical assumptions regarding the data collection process and the structure of the variables. These prerequisites must be met to ensure that the resulting p-value and statistical conclusion are accurate and reliable. Violating these assumptions can lead to skewed results, erroneous interpretation, and ultimately, invalid conclusions drawn from the research. Understanding and verifying these conditions is an essential step before proceeding with the analysis.

Furthermore, while the McNemar Test is generally considered non-parametric, meaning it does not require the assumption of normality, it does rely heavily on the structure of the paired observations and the underlying distribution being suitable for the chi-square approximation. A key practical assumption often cited is the requirement for a sufficient number of discordant pairs. Although the official prerequisites focus on data collection, practitioners must ensure that the number of cells B and C (the discordant outcomes) is large enough (often suggesting $B+C \geq 10$) to justify the use of the standard large-sample chi-square approximation.

The fundamental assumptions that underpin the statistical integrity of the McNemar Test include:

Random Sample Selection

Dependent/Paired Observations

Dichotomous Response Variables (Mutually Exclusive Groups)

The Necessity of Random Sampling: Ensuring Data Integrity

The first and most universal assumption in statistical inference is that the data points for the groups under analysis must originate from a simple random sample drawn from the target population. This means every individual in the population had an equal chance of being selected for the study. Adherence to strict random sampling procedures is paramount because it ensures that the sample is representative of the larger population we wish to generalize our findings to. If sampling is non-random, the resulting data is inherently subject to bias.

When selection bias is present--for instance, if specific demographic groups are over-represented or systematically excluded--the statistical analysis, regardless of its computational accuracy, will produce results that are inaccurate representations of the population parameters. This systematic error, or **bias**, invalidates the use of inferential tests like McNemar's because the foundation upon which statistical significance rests--the random probability of observing the data--is compromised. Therefore, researchers must meticulously document their sampling methodology to ensure this assumption is satisfied.

The Nature of Dependent or Paired Samples

A core distinguishing feature of the McNemar Test is its reliance on **paired samples**, meaning that the two observations being compared are fundamentally linked or dependent. This dependency structure usually arises in one of two ways: either the same group of subjects is observed repeatedly across two different time points (a repeated measures design), or two different subjects are matched based on specific criteria (a matched-pair design). For example, if a group of patients is assessed for disease status (Yes/No) before and after receiving a novel therapy, the two observations for each patient constitute a paired sample.

The dependence is critical because the test's calculation specifically leverages the relationships within these pairs, focusing only on the shifts in outcome. Unlike tests for independent samples (like the standard Chi-Square Test of Independence), the McNemar Test removes inter-subject variability by using each subject as their own control. If the samples were mistakenly treated as independent, the statistical power would be significantly reduced, and the resulting inference regarding the change in proportions would be incorrect. This paired structure is what gives the McNemar Test its statistical advantage in longitudinal or crossover studies.

Defining Mutually Exclusive Outcomes (Dichotomous Variables)

The response variable of interest in the McNemar Test must be **dichotomous**, meaning it must have exactly two possible outcomes. Furthermore, these outcomes must be **mutually exclusive**. A subject cannot simultaneously belong to both categories at the same measurement point. Classic examples of appropriate dichotomous variables include success/failure, recovered/not recovered, or vote/did not vote. This structure is what allows the data to be placed neatly into the 2x2 contingency table that the McNemar Test analyzes.

For instance, if your categorical variable is 'Hungry' (Yes/No), a single person cannot report being both 'Yes' and 'No' simultaneously. If the variable had more than two categories (e.g., Small, Medium, Large), the McNemar Test would be inappropriate, and alternative non-parametric tests like the Cochran's Q test might be required if the data is still paired and nominal. Ensuring the categorical variable is binary and mutually exclusive is foundational to correctly calculating the test statistic based on the chi-square distribution.

Determining the Appropriate Use Case for the McNemar Test

Identifying the correct statistical test is paramount to sound research methodology. The McNemar Test is highly specialized, making its application appropriate only under specific data and research design conditions. Researchers should conduct a careful review of their variables and sampling structure before selecting this test. If the data structure deviates from these key requirements, a different analytical approach, such as the Paired Samples T-Test (for continuous data) or the Chi-

Square Test of Independence (for independent groups), would be necessary.

The ideal scenario for deploying the McNemar Test is rooted in studies that employ repeated measures or matched pairs where the outcome of interest is binary. It is the go-to technique for investigating change over time or comparing two related diagnostic procedures. Before proceeding, verify that your scenario satisfies the following five crucial criteria:

The objective is to test for a **difference** in outcomes between the two related measurements.

The variable of interest must be fundamentally **categorical** or derived as a proportion.

The categorical variable must possess exactly **two options** (dichotomous).

The data must originate from **paired samples** or dependent groups.

For the standard asymptotic test, the cell counts of the **discordant pairs** (cells B and C) should ideally sum to more than 10.

Focusing on Differences, Not Relationships or Prediction

The purpose of the McNemar Test is highly specific: it is designed to examine whether a significant **difference** exists between the two measurements taken under paired conditions. This is distinct from other analytical goals, such as testing for an association or relationship between two variables (which might use a simple correlation or Chi-Square Test of Independence), or predicting the value of one variable based on another (which involves regression analysis). The research question must specifically involve quantifying the change or effect of an intervention when comparing Condition 1 to Condition 2 on the same subjects.

If your hypothesis, for instance, is that a new drug increases the proportion of patients recovering compared to the baseline measurement, you are explicitly testing for a difference in recovery rates. This directional focus ensures that the McNemar Test is addressing the appropriate statistical query. If the research were instead focused on whether Recovery Status is related to Age (an independent variable), the McNemar Test would be entirely unsuitable, as it requires both variables being compared to be the paired measurement of the same dichotomous outcome.

Handling Proportional and Categorical Data Structures

The dependent variable analyzed by the McNemar Test must be either intrinsically categorical or represented by proportions derived from categorical counts. A **categorical variable** is defined as a variable whose values represent groups or categories without a natural numerical order; classic examples include eye color (blue, green, brown), type of consumer response (click, ignore), or diagnostic outcome (positive, negative). The data input into the McNemar table consists of counts of individuals falling into these categories across the two time points.

Proportional variables are merely the calculated ratios or percentages of these categorical counts.

For instance, comparing the percentage of customers who converted on Website Version A (10%) versus Website Version B (15%) requires the underlying data to be the count of conversions (Yes/No), which is categorical. This relationship between counts and proportions is why the test is often described as comparing "paired proportions."

*It is vital to reiterate that if the data points are **continuous variables**--such as blood pressure, weight, or IQ score--the McNemar Test is inappropriate. Researchers should instead utilize methods designed for metric data, such as the Paired Samples T-Test, which compares the means of the two related groups.*

The Requirement for Dichotomous Outcomes

For the McNemar Test to function correctly, the categorical variable must be reduced to exactly two mutually exclusive options, often referred to as a binary or **dichotomous variable**. While some variables are naturally binary (e.g., gender, recovery status: Yes/No), others with multiple categories may need to be collapsed into two groups (e.g., grouping 'low' and 'medium' satisfaction together versus 'high' satisfaction). This requirement facilitates the construction of the essential 2x2 contingency table needed for the calculation.

Examples of appropriate dichotomous variables include whether a respondent made a purchase (Yes/No), the outcome of a medical screening (Positive/Negative), or whether a belief shifted after an intervention (Agree/Disagree). If your variable retains three or more options (e.g., Excellent, Good, Poor), you cannot directly apply the McNemar Test; you must either collapse the categories or select an alternative test for multiple paired nominal categories, such as Cochran's Q test, which is an extension of McNemar's for more than two related groups.

The Necessity of Paired Data Structures

As highlighted in the assumptions section, the requirement for **paired samples** is non-negotiable for the McNemar Test. Paired data implies a structural link between the two observations. The most common structural links are **repeated measures** (Time 1 vs. Time 2 on the same subjects) or **matched subjects** (e.g., treating one member of a pair with Drug A and the other with Drug B, where pairs are matched based on demographics or disease severity). This structure is essential because the McNemar Test specifically computes the difference by comparing the changes within each pair, isolating the effect of the intervention or time.

A clear example of paired data involves tracking a cohort of employees before and after a training program to see if their performance rating shifts from 'Unsatisfactory' to 'Satisfactory'. The two performance ratings are dependent measurements. If, however, you were comparing the recovery rate of one group of patients who received Drug A to an entirely separate, unmatched group of patients who received Drug B, the samples would be independent, and you would need to use a

standard Chi-Square Test of Independence instead of the McNemar Test.

Practical Application and Interpretation: A Disease Recovery Example

To illustrate the utility of the McNemar Test, consider a common scenario in clinical research: evaluating the impact of a targeted intervention on disease status. Imagine a study where researchers administer a new drug treatment to a cohort of patients suffering from a chronic condition. The primary measure of interest is whether the patient has **Recovered from Disease** (a dichotomous outcome: Yes/No). This status is measured at two distinct time points: **Baseline** (Time 1, before treatment) and **Follow-up** (Time 2, after treatment).

The data is structured into a 2x2 table where the rows represent the outcome at Time 1 and the columns represent the outcome at Time 2. The critical cells for the McNemar calculation are the discordant cells: Cell B (recovered at T1, not recovered at T2) and Cell C (not recovered at T1, recovered at T2). We are fundamentally interested in whether the number of people who improved (Cell C) is significantly different from the number of people who worsened (Cell B). This design perfectly satisfies the conditions for the McNemar Test: the variables are categorical (Yes/No), the groups are paired (repeated measures), and the goal is to assess a difference.

In this specific context, the formal statistical hypotheses are established as follows: The **Null Hypothesis (H₀)** posits that there is no true difference in the recovery rates between the two measurements (i.e., the proportion of shifts from Yes to No equals the proportion of shifts from No to Yes). Conversely, the **Alternative Hypothesis (H_a)** asserts that a statistically significant difference exists in the recovery proportions. The test then determines whether the observed changes--the discordant pairs--are sufficiently uneven to reject the notion that the shift happened purely by chance.

Interpreting the P-Value and Drawing Conclusions

The output of the McNemar Test is a test statistic, which is based on the calculation of $\frac{(\text{Cell B} - \text{Cell C})^2}{(\text{Cell B} + \text{Cell C})}$ and is then used to calculate a corresponding **p-value**. The p-value is perhaps the most crucial output, representing the probability of observing the detected differences in recovery rates (or an even more extreme difference) if the null hypothesis were actually true. In simpler terms, it quantifies the risk of concluding that a difference exists when, in reality, there is none. Lower p-values indicate that the observed data is inconsistent with the null hypothesis.

Researchers typically compare the calculated p-value to a predetermined threshold, known as the significance level (α), which is conventionally set at 0.05. If the resulting p-value is less than or equal to 0.05 ($p \leq 0.05$), the result is deemed statistically significant. This outcome leads to the rejection of the null hypothesis, allowing the researcher to conclude with confidence that the

change in recovery rate between Time 1 and Time 2 is real and not merely due to random fluctuation or chance. For example, if $p = 0.01$, there is only a 1% chance of seeing the observed shift if the treatment had absolutely no effect.

Conversely, if the p-value is greater than 0.05, the results are considered non-significant. In this case, the researcher fails to reject the null hypothesis, concluding that there is insufficient evidence to claim that the treatment caused a meaningful change in the recovery proportions. It is essential for rigorous reporting that the exact p-value, the test statistic, and the sample size (especially the count of discordant pairs) are documented to allow for full transparency and replicability of the findings. The McNemar Test provides a clear, precise measure of directional change in binary outcomes for dependent data.

ARABPSYCHOLOGY.COM