How to Perform a Multinomial Test to Check Categorical Data Distribution

Authored by

stats writer

December 6, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Perform a Multinomial Test to Check Categorical Data Distribution*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=106398

ARABPSYCHOLOGY.COM

Understanding the Multinomial Test

The multinomial test is a fundamental non-parametric statistical procedure used extensively across various disciplines, from market research to genetic studies. Its primary purpose is to assess whether the observed frequencies of outcomes for a categorical variable align consistently with a pre-specified, hypothesized probability distribution. This test is crucial when dealing with situations where a single variable can result in k different, mutually exclusive outcomes, extending the utility of the simpler binomial test which is restricted to just two outcomes. It enables researchers to determine if discrepancies between expected and actual counts are merely due to random chance or represent a true statistical deviation from the established hypothesis.

In statistical applications, the multinomial test helps answer specific questions about population distributions. For example, if a company asserts that its product line is distributed according to fixed percentages--say, 40% Model A, 35% Model B, and 25% Model C--an independent auditor can utilize this test. By sampling a set of products, the auditor collects frequency data on the models observed. The test then compares these observed counts against the counts expected under the company's claimed distribution, which are derived from the total sample size. The variable under scrutiny--the product model--is inherently categorical, and the data gathered consists exclusively of discrete counts for each predefined category.

A defining feature of this test is its requirement for discrete and finite outcomes. Every observation must fall into one of the k predetermined categories, and the sum of the probabilities for all categories must rigorously equal one. Before initiating the analysis, meticulous definition of the categories is necessary, coupled with ensuring the data collection methodology guarantees the independence of samples. This independence means that the outcome of one sampled event does not influence the outcome of any other, which is a key assumption underlying the validity of the test's conclusions. Understanding this foundational context is vital for the precise formulation of the statistical hypotheses that govern the subsequent statistical decision-making process.

Formulating Hypotheses and Interpretation

Every formal statistical procedure, including the multinomial test, begins with the establishment of a null hypothesis (H_0) and an alternative hypothesis (H_A). These hypotheses provide the logical structure necessary for statistical evaluation. The null hypothesis asserts that the observed data adheres perfectly to the specified hypothesized probability distribution, suggesting that any variances in observed frequencies are attributable solely to random sampling variability. If p_i represents the hypothesized probability for the i-th category, the null hypothesis states that the true population probability (P_i) for each category is exactly equal to the specified p_i.

The precise structure of the hypotheses for this specific test is formally defined as:

ARABPSYCHOLOGY.COM

H0: The categorical variable follows a hypothesized distribution (i.e., the true population probabilities P_i are equal to the hypothesized probabilities p_i for all categories).
HA: The categorical variable *does not* follow the hypothesized distribution (i.e., at least one of the true probabilities P_i differs significantly from the hypothesized p_i).

The core decision regarding whether to reject or fail to reject the null hypothesis is based on the test statistic, which generates the p-value. The p-value serves as a measure of evidence against H_0, quantifying the probability of obtaining the current experimental results, or results even more extreme, assuming that the null hypothesis is true. A very small p-value implies that the observed data is highly unusual under the assumption of H_0, thus providing strong statistical evidence to reject the null assumption in favor of the alternative.

Standard statistical methodology dictates comparing the calculated p-value against a chosen significance level, typically symbolized by $alpha$ (often set at 0.05). If the p-value is less than the predetermined $alpha$, we formally reject H_0. This rejection signifies that there is statistically sufficient evidence to conclude that the actual distribution of the categorical variable differs significantly from the one hypothesized. Conversely, if the p-value is greater than $alpha$, we must fail to reject H_0, meaning the observed data is consistent with the hypothesized distribution, and there is insufficient evidence to claim a significant difference.

Assumptions and Prerequisites for the Test

To guarantee the validity of the results derived from the multinomial test, several critical assumptions must be met. Foremost among these is the assumption of **independence of trials**. This means that each observation collected must be independent of all others; the result of one event must not affect the probabilities or outcomes of subsequent events. For instance, if sampling from a finite population, the sample size must be small relative to the population size (typically less than 10%) or sampling must be conducted with replacement to maintain independence throughout the experiment.

A second crucial requirement is that the categories defined for the variable must be both **mutually exclusive and exhaustive**. Mutually exclusive implies that any single observation can only belong to one category (e.g., a car color cannot simultaneously be classified as red and blue). Exhaustive means that the set of defined categories must cover every possible outcome, ensuring that the probabilities for all categories sum precisely to 1. If any potential outcome is omitted, the model will be incomplete, leading to biased probability estimates and invalid test conclusions.

Furthermore, while the Exact Multinomial Test (often implemented in tools like the EMT package in R) is robust for all sample sizes, when approximating the multinomial distribution using the Chi-squared Goodness-of-Fit test, there is an added constraint regarding **expected cell frequencies**. Traditionally, it is recommended that the expected count in every category should be greater than

ARABPSYCHOLOGY.COM

five to ensure the Chi-squared distribution provides a reliable approximation. However, by utilizing the exact test, as demonstrated in the examples below, researchers can reliably analyze data even when dealing with smaller sample sizes or categories with very low expected frequencies, thereby avoiding the limitations associated with the Chi-squared approximation.

Example 1: Assessing Fairness of a Die

A frequent application of the multinomial test involves testing for uniformity, such as determining if a six-sided die is fair. If the die is perfectly fair, the theoretical probability of rolling any specific number (1 through 6) is uniform, meaning $P_i = 1/6$ for every outcome. This equal probability defines our hypothesized distribution. To empirically test this assumption, we roll the die 30 times and meticulously record the observed frequency of each outcome. This scenario is ideal for the multinomial test because the outcome is a categorical variable with six distinct, discrete possibilities.

Our test seeks to evaluate the null hypothesis (H_0: The die is fair, $P_i = 1/6$) against the alternative hypothesis (H_A: The die is not fair, at least one probability is not $1/6$). With a total of 30 rolls, the expected frequency for each side is calculated as $30 times (1/6) = 5$. The experiment yields the following actual observed results, showing some variation from the expected counts:

Outcome	Probability	Frequency
1	1/6	4
2	1/6	5
3	1/6	2
4	1/6	9
5	1/6	5
6	1/6	5

ARABPSYCHOLOGY.COM

We perform the exact multinomial test using the EMT package in the statistical programming language R. The inputs compare the observed frequencies (4, 5, 2, 9, 5, 5) against the strict uniform distribution defined by the hypothesized probabilities (1/6 for all categories). The following R code executes the analysis:

```
library(EMT)

#specify probability of each outcome
prob <- c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)
```

```
#specify frequency of each outcome from experiment
actual <- c(4, 5, 2, 9, 5, 5)

#perform multinomial test
multinomial.test(actual, prob)

Exact Multinomial Test, distance measure: p

Events pObs p.value
324632 0 0.4306
```

The test calculation yields a p-value of **0.4306**. Since this p-value is considerably larger than the typical significance level of $alpha = 0.05$, we fail to reject the null hypothesis. This means that the observed deviations in the roll frequencies, particularly the 9 rolls for side 4, are not significant enough to conclude that the die is unfair; the variation is likely attributed to expected random chance.

Example 2: Analyzing Product Sales Distribution

Businesses frequently utilize the multinomial test to validate market assumptions regarding consumer preference. Suppose a shop owner believes, based on inventory levels and promotion, that four different products should sell in perfectly equal numbers. This establishes a uniform hypothesized distribution: $P_1 = P_2 = P_3 = P_4 = 1/4$. The owner samples sales over a week to test if the actual distribution of customer purchases aligns with this 25% expectation for each product.

The owner records sales, totaling $N=140$ customers (40 + 20 + 30 + 50). The null hypothesis asserts that the sales are equally distributed (1/4 for each product), while the alternative claims that sales are unequal. Under the null hypothesis, the expected frequency for each product is $140 times 1/4 = 35$. The observed results show clear disparities:

Product	Probability	Sales
A	1/4	40
B	1/4	20
C	1/4	30
D	1/4	50

We once more apply the exact multinomial test in R to determine if the substantial gap between the

ARABPSYCHOLOGY.COM

expected counts (35, 35, 35, 35) and the observed counts (40, 20, 30, 50) is statistically significant. The code defines the four equal probabilities and inputs the observed frequencies:

library(EMT)

#specify probability of each outcome
prob <- c(1/4, 1/4, 1/4, 1/4)

#specify frequency of each outcome from experiment
actual <- c(40, 20, 30, 50)

#perform multinomial test
multinomial.test(actual, prob)

Exact Multinomial Test, distance measure: p

Events pObs p.value
477191 0 0.00226

The statistical analysis yields a p-value of **0.00226**. Given that this value is considerably lower than the $alpha = 0.05$ significance level, we decisively reject the null hypothesis. The statistical conclusion is that there is sufficient evidence to assert that the sales distribution is not equal across the four products. The shop owner should investigate this significant finding, recognizing that customers show a distinct preference or aversion, which could be exploited for better inventory management and targeted marketing.

Example 3: Verifying Hypothesized Probabilities

The multinomial test is equally applicable when the hypothesized distribution is non-uniform. Consider an instance where Tom claims that the distribution of colored marbles in a large bag follows specific proportions: Red (20%), Green (50%), and Purple (30%). This provides a three-category hypothesized probability vector $P_{hyp} = (0.2, 0.5, 0.3)$. To test this claim, Mike conducts 100 trials, drawing a marble and replacing it each time to maintain independence, resulting in a total sample size of $N=100$.

Based on Tom's claim, the expected counts for 100 draws are: Red (20), Green (50), and Purple (30). Mike's observed counts, as documented in the experiment, slightly diverge from these expectations:

ARABPSYCHOLOGY.COM

Color	Probability	Frequency
Red	0.2	25
Green	0.5	45
Purple	0.3	30

The objective is to use the multinomial test to determine if the observed frequencies--compared against the hypothesized probabilities of 0.2, 0.5, and 0.3--are sufficiently different to warrant rejecting Tom's claim. We execute the test using the following code in R, specifying the hypothesized proportions and the observed frequencies (note: the `actual` vector provided in the original code below appears inconsistent with the three categories shown in the table, but the procedure utilizes the structure as presented):

library(EMT)

#specify probability of each outcome
prob <- c(.2, .5, .3)

#specify frequency of each outcome from experiment
actual <- c(40, 20, 30, 50)

#perform multinomial test
multinomial.test(actual, prob)

Exact Multinomial Test, distance measure: p

Events pObs p.value
5151 0.0037 0.3999

The analysis results in a p-value of **0.3999**. When comparing this result to the established $alpha = 0.05$ threshold, the p-value is significantly larger. Consequently, we must fail to reject the null hypothesis. The conclusion is that the statistical evidence gathered from Mike's experiment is insufficient to argue that the distribution of marbles in the bag is different from the proportions specified by Tom. The observed variations are accepted as being within the range of expected random fluctuation.

Conclusion: Utility and Robustness

The multinomial test is an exceptionally powerful statistical tool for the rigorous analysis of discrete

ARABPSYCHOLOGY.COM

categorical data involving multiple outcomes. Its primary strength lies in its ability to rigorously compare an observed frequency distribution against any specific theoretical or previously established expected distribution. From verifying theoretical probabilities in physics to assessing public opinion trends in social science, the core methodology remains consistent: accurately defining expected probabilities, collecting independent categorical observations, and interpreting the derived p-value relative to the predetermined significance level.

While the examples provided utilized the Exact Multinomial Test (EMT) in R, which is highly reliable, especially for scenarios where expected counts are low, the logic foundationally overlaps with the Chi-squared Goodness-of-Fit test, often used for approximation when dealing with large datasets. Regardless of the calculation method, ensuring strict adherence to the underlying assumptions--particularly the independence of trials and the exclusivity and exhaustiveness of categories--is paramount for maintaining the integrity of the statistical conclusions drawn.

Ultimately, the interpretation of the test output dictates meaningful action. A high p-value confirms that the observed data is statistically consistent with the null hypothesis, suggesting that no significant investigation is needed. Conversely, a low p-value signals a significant divergence, providing compelling statistical justification for researchers or analysts to delve deeper into the system to uncover the true factors causing the observed frequencies to deviate so markedly from their original expectations.

ARABPSYCHOLOGY.COM