Calculating Odds Ratios with SAS: A Tutorial for Statistical Analysis

Authored by
Mohammed looti

November 14, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Calculating Odds Ratios with SAS: A Tutorial for Statistical Analysis*. PSYCHOLOGICAL STATISTICS. Retrieved from https://statistics.arabpsychology.com/?p=1274

In the rigorous world of quantitative research and statistics, researchers are frequently tasked with quantifying the relationship between distinct factors, especially when the outcome of interest is binary (e.g., success/failure, presence/absence). The Odds Ratio (OR) stands out as one of the most powerful and broadly utilized metrics for this purpose, particularly within observational study designs. Calculating the OR provides essential, evidence-based insights into the relative likelihood of a specific event occurring in one defined group compared to another, forming a cornerstone for rigorous decision-making across diverse scientific and clinical disciplines.

The core function of the **odds ratio** is to precisely quantify the ratio of the odds of an outcome happening within an exposed group relative to the odds of that same outcome happening within an unexposed group. Its utility spans vast domains, from epidemiology and public health policy development to clinical trials and sociological research. In these fields, accurately assessing the magnitude and strength of the association between a predictor variable (the exposure) and a result (the outcome) is absolutely paramount for drawing valid conclusions.

To accurately calculate and rigorously interpret this measure, raw data must first be organized, typically structured within a 2-by-2 contingency table. This tabular format efficiently cross-tabulates the frequencies of two distinct categorical variables, each defined by two specific levels. For the robust calculation and comprehensive inferential analysis of these frequency counts, the use of sophisticated statistical software is indispensable. The SAS (Statistical Analysis System) software suite is a world leader in this area, offering specialized procedures explicitly designed for advanced categorical data analysis.

	Event	No Event
Treatment	A	B
Control	C	D

Within the SAS environment, the **PROC FREQ** statement functions as the primary utility for efficiently computing the odds ratio, alongside various related measures of association and essential statistical tests tailored for frequency data. This comprehensive guide is designed to provide you with a clear, step-by-step methodology, covering everything from initial data setup and implementing the necessary SAS syntax to the final critical step of interpreting the output to derive statistically meaningful conclusions about the relationship observed between your variables.

Quantifying Association: Defining the Odds Ratio

The odds ratio is fundamentally conceptualized as a measure of the multiplicative change in the odds of observing an outcome that is directly associated with a shift in the exposure status. This metric is absolutely crucial for researchers seeking to understand and quantify risk factors, protective factors, or treatment effects within binary outcome scenarios. Unlike simpler descriptive statistics, the OR provides a single, universally interpretable number that concisely summarizes the potential impact or association stemming from an exposure.

The interpretation of the OR is intuitive and highly structured: an odds ratio that calculates out to exactly 1.0 serves as the null value, indicating that there is absolutely no statistical association between the exposure and the outcome; the odds of the event are precisely the same in both the exposed and unexposed groups. Conversely, an odds ratio significantly greater than 1.0 suggests a positive association, implying that the exposure effectively increases the odds of the outcome occurring. Conversely, an odds ratio less than 1.0 indicates a negative association, where the exposure is linked to substantially reduced odds of the outcome. For practical illustration, an OR of 2.5 means that the odds of the outcome occurring are 2.5 times higher in the exposed group when compared to the unexposed group.

The mathematical calculation of the OR is derived directly from the cell counts recorded within the 2-by-2 contingency table. If we adopt standard notation where the cell counts are denoted as A (exposed cases), B (exposed non-cases), C (unexposed cases), and D (unexposed non-cases), the formula simplifies elegantly to **(A × D) / (B × C)**. This powerful calculation represents the ratio of the odds of the outcome in the exposed group (calculated as A/B) divided by the odds of the outcome in the unexposed group (calculated as C/D).

It is critically important for analysts to maintain a clear conceptual distinction between the odds ratio and the relative risk (RR). While both are measures designed to assess association, the RR is defined as a ratio of probabilities (or risks), whereas the OR is strictly defined as a ratio of odds. Although these two measures can yield numerically similar results, especially in situations where the outcome is statistically rare (prevalence typically below 10%), they represent conceptually distinct interpretations of effect size. Furthermore, in specific study designs, such as cross-sectional studies or case-control studies, the OR often represents the only feasible and valid estimate of the underlying association.

Leveraging SAS for Categorical Data: The PROC FREQ Procedure

The **PROC FREQ** procedure is often considered the essential workhorse of categorical data analysis within the SAS statistical environment. This procedure is exceptionally robust and versatile, designed to efficiently generate not only simple one-way frequency distributions but also to construct the complex n-way frequency tables that are necessary for meticulously assessing multivariate relationships and associations. For the specific analysis of 2-by-2 tables, **PROC FREQ** grants immediate and reliable access to all necessary inferential statistics, including the OR.

When analyzing any contingency table, **PROC FREQ** requires the user to clearly specify which variables serve as the row and column variables by utilizing the mandatory `TABLES` statement. To explicitly request the odds ratio and other associated measures of risk (like relative risk), the analyst must include the `RELRISK` option in this statement. Additionally, it is standard practice to incorporate the `CHISQ` option, as this provides various critical chi-square tests that assess the statistical independence of the row and column variables, results which are frequently reported in conjunction with the calculated odds ratio.

A significant practical advantage of **PROC FREQ** is its inherent flexibility in managing diverse types of input data. The procedure is fully equipped to analyze raw data, where every row represents a single observation, but it is equally efficient--and often preferred--when dealing with summary data where each row represents an aggregated count for a specific combination of categories. In the latter scenario, the use of the `WEIGHT` statement becomes absolutely indispensable; it ensures that the provided counts are correctly incorporated into the frequency analysis, guaranteeing the derivation of accurate and statistically representative results.

Preparing Data for Analysis: A Practical Basketball Example

To provide a clear, practical demonstration of calculating the odds ratio using SAS, we will establish a hypothetical scenario focused on performance evaluation in sports. Imagine a controlled study designed specifically to compare the effectiveness of two distinct training programs: a New Program (designated as the exposure group) versus an Old Program (designated as the unexposed group). The outcome measure is the result of a standardized skills test (binary outcome: Pass/Fail). The study involves 100 basketball players, with 50 allocated randomly to each of the two programs.

The primary research objective is to quantitatively measure the difference in the odds of successfully passing the skills test based solely on the training program utilized. To facilitate the accurate OR calculation, the collected results must be meticulously aggregated into the standard 2-by-2 format, ensuring a clear delineation between the exposure group (New Program vs. Old Program) and the outcome variable (Passed vs. Failed). This structured organization is prerequisite for the **PROC FREQ** procedure to correctly identify and assign the cell counts (A, B, C, D) required by the core odds ratio formula.

The resulting hypothetical frequency outcomes derived from this structured study are presented below. This table clearly displays the counts for every combination of the training program and the skills test result:

	Passed	Failed
New Program	34	16
Old Program	39	11

Our immediate goal is to translate this summarized count data efficiently into a SAS dataset and then execute the appropriate statistical analysis to derive the final odds ratio. This calculated OR will serve as the crucial quantitative measure, allowing us to compare the odds of passing the test for players who participated in the New Program relative to those who used the Old Program, thereby assessing the comparative effectiveness of the new training regimen.

Executing the Calculation: Implementing PROC FREQ Syntax

The technical process of calculating the odds ratio within SAS commences with the precise definition of the dataset containing the structured frequency counts. Since our example data is already summarized, we strategically employ the `DATALINES` statement to input the three necessary variables: the test result, the training program, and the aggregated count corresponding to that combination.

The subsequent **PROC FREQ** step requires the careful application of several key options to ensure correct analysis. The `WEIGHT` statement is absolutely critical in this context; it explicitly instructs SAS that the variable named `count` represents the frequency of observations in that row, rather than treating each row of input as a single observation unit. The `TABLES` statement is then used to specify the exact structure of the desired 2-by-2 table, defining `program` as the row variable and `result` as the column variable. Finally, the inclusion of both the `CHISQ` and `RELRISK` options ensures that the output will contain all the essential statistical measures required for robust inference and accurate association measurement.

The following syntax demonstrates the complete, executable process required to generate the OR:

```
/*create dataset*/
data my_data;
input result $ program $ count;
datalines;
Passed New 34
Passed Old 39
_Failed New 16
_Failed Old 11
;
```

```
run;
/*calculate odds ratio*/
proc freq data=my_data;
weight count;
tables program * result / chisq relrisk;
run;
```

Executing this script first successfully creates the necessary `my_data` dataset, guaranteeing that the counts are accurately associated with their respective categorical levels. Following this, the **PROC FREQ** procedure processes this weighted data to generate the required frequency tables and the detailed measures of association, preparing the results for subsequent, critical interpretation.

Analyzing the Results: Deciphering SAS Output Tables

Upon the successful execution of the SAS code, the output generated by **PROC FREQ** delivers a comprehensive set of tables detailing the statistical analysis. The initial table presented is the frequency distribution itself, which serves as a direct, visual representation of the 2-by-2 contingency table structure. This foundational table is vital for initial data verification, as it displays the observed counts, overall percentages, row percentages, and column percentages for every cell combination.

A careful review of this initial frequency table confirms that SAS has correctly mapped the row variable (Program) and the column variable (Result) and accurately applied the frequency counts specified in the input data. Achieving this visual confirmation is a crucial prerequisite step before proceeding to interpret any of the derived inferential statistics.

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of program by result			
	result		
program	Passed	_Failed	Total
New	34 34.00 68.00 46.58	16 16.00 32.00 59.26	50 50.00
Old	39 39.00 78.00 53.42	11 11.00 22.00 40.74	50 50.00
Total	73 73.00	27 27.00	100 100.00

Following the presentation of these descriptive statistics, SAS provides specialized tables related to statistical testing and measures of association. The most critical output for our immediate research objective is clearly labeled "Odds Ratio and Relative Risk Estimates." This dedicated section directly furnishes the calculated point estimate for the odds ratio, along with its associated confidence interval, which is essential for assessing precision.

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	0.5994	0.2449	1.4666
Relative Risk (Column 1)	0.8718	0.6855	1.1088
Relative Risk (Column 2)	1.4545	0.7518	2.8144

Sample Size = 100

Based on the output generated by the `RELRISK` option, the calculated odds ratio for this specific study is determined to be **0.5994**. This value quantitatively summarizes the relationship between the type of training program and the likelihood of a passing outcome. Since this OR value is less than 1.0, it directly suggests a negative association: the odds of successfully passing the skills test are lower for players who utilized the New Program compared to those who utilized the Old Program. Specifically, the odds of success are approximately 40.06% lower (derived from 1 -

0.5994) when the new training regimen is implemented, indicating that the baseline Old Program was comparatively more effective in this sampled population.

Assessing Reliability: Statistical Significance and Confidence Intervals

While the point estimate (OR = 0.5994) provides a precise measure of the observed association within our specific sample, it is equally vital to determine the overall reliability and generalizability of this finding to the broader population. This critical assessment is primarily conducted through the evaluation of the confidence interval (CI), which is readily provided by the `RELRISK` option within **PROC FREQ**.

For this basketball example, the 95% confidence interval reported for the odds ratio is . This interval carries significant meaning: it signifies that if we were to theoretically replicate this study numerous times, 95% of the resulting confidence intervals would successfully capture the true population odds ratio between the New Program (exposure) and the passing outcome relative to the Old Program (reference). The CI thus effectively provides a range of highly plausible values for the true population effect, while explicitly accounting for natural sampling variability and uncertainty.

The definitive statistical test for statistical significance at the standard 0.05 alpha level (corresponding to a 95% CI) centers on whether the calculated interval encompasses the null value of 1.0. An odds ratio of exactly 1.0 represents the null hypothesis--that is, there is zero association or no difference in odds between the two training programs. If the 95% confidence interval includes the value 1, we must fail to reject the null hypothesis; this implies that the observed difference is not deemed statistically significant and could reasonably be attributed to random chance or sampling error.

Crucially, in our basketball example, the confidence interval clearly contains the value of 1.0. Therefore, despite the point estimate (0.5994) suggesting substantially lower odds for the New Program, our statistical conclusion is that this observed difference is not **statistically significant**. We do not possess sufficient statistical evidence to confidently conclude that the New Program truly alters the odds of success compared to the Old Program in the entire population of basketball players.

Conclusion and Next Steps in Categorical Data Analysis

A proficient grasp of both the calculation and the nuanced interpretation of the odds ratio is an essential skill for any quantitative researcher dealing with binary outcomes. By strategically leveraging highly capable statistical tools like SAS and the robust **PROC FREQ** procedure, researchers can efficiently and accurately transform raw frequency data into highly informative measures of association and reliability.

While **PROC FREQ** is exceptionally effective and sufficient for the analysis of simple 2-by-2 tables, more sophisticated research questions involving multiple potential confounding variables often necessitate the use of advanced statistical techniques. This typically involves employing logistic regression (often using **PROC LOGISTIC** in SAS), which is capable of calculating adjusted odds ratios that control for multiple factors simultaneously. Nevertheless, a comprehensive understanding of the foundational principles taught using **PROC FREQ** provides the necessary strong foundation for seamlessly transitioning to these more complex statistical modeling approaches.

To further deepen your expertise in effectively handling categorical data and rigorously interpreting measures of association, we recommend exploring the following authoritative resources:

Official SAS Documentation for `PROC FREQ`: This resource provides the most comprehensive and authoritative details regarding all available options, underlying statistical assumptions, and overall functional capabilities of the procedure.
Advanced Biostatistics and Epidemiology Textbooks: These offer detailed theoretical background on the OR, various study designs, and the critical distinction between the OR and relative risk.
Statistical Inference Guides: Consulting these guides will enhance your understanding of formal hypothesis testing, the proper application of P-values, and the correct interpretation of the confidence interval, especially in the context of null effects.