How to Perform a Repeated Measures ANOVA in Python

Authored by

stats writer

December 25, 2025

A Repeated Measures ANOVA (Analysis of Variance) represents a robust statistical technique essential for analyzing data derived from within-subjects experimental designs. This method is specifically employed to determine if there is a statistically significant difference among the means of three or more related treatment conditions, where the same group of subjects is exposed to every condition. Unlike a traditional one-way ANOVA, the repeated measures design accounts for the correlation between measurements taken from the same individual, enhancing statistical power and controlling for inherent subject variability. Understanding its application is crucial for researchers utilizing longitudinal studies or crossover designs, ensuring that the appropriate statistical model is applied to correlated data.

This comprehensive tutorial provides an expert guide on how to execute a one-way repeated measures ANOVA using the powerful statistical capabilities available within the Python programming environment. We will utilize established libraries such as Pandas for efficient data management and the specialized statsmodels library for the rigorous statistical analysis itself, ensuring reproducible and accurate results. The focus here is on the practical implementation and subsequent interpretation of the ANOVA output, moving beyond theoretical assumptions to actionable insights.

The Statistical Need for Repeated Measures Design

The core distinction of a repeated measures design is that the independent variable is manipulated within the same subjects. This is highly advantageous in fields like psychology or pharmacology because it drastically reduces the influence of confounding variables related to individual differences, such as genetic predisposition, baseline cognitive ability, or general physiological reactivity. By using the same individuals across all conditions, we effectively use each subject as their own control, leading to a much cleaner comparison between the treatment levels. This design principle is fundamental to achieving high internal validity in experimental settings where minimizing noise is paramount.

When analyzing such data, standard ANOVA techniques are insufficient because they assume independence between observations. Since measurements taken from the same patient across different drugs are inherently correlated, ignoring this dependency violates a core assumption of independent samples tests. The repeated measures ANOVA corrects for this by partitioning the total variance into three components: variance due to the treatment (the within-factor effect), variance due to individual differences (the subject effect), and residual error. This partitioning allows us to isolate the true effect of the treatment with greater precision than non-repeated measures alternatives.

Furthermore, a crucial assumption underlying the validity of the F-test in repeated measures ANOVA is sphericity. Sphericity refers to the condition where the variances of the differences

between all pairs of within-subject conditions are equal. Violations of sphericity can lead to an inflated F-statistic and an increased risk of Type I error. While the Python `statsmodels` implementation often provides methods to address or test for this assumption in more complex multivariate designs, for a simple one-way analysis, we focus on correctly specifying the subject and within-group factors to ensure the variance calculation is appropriate for the correlated structure of the data.

Example Scenario: Investigating Drug Efficacy on Reaction Time

To illustrate the application of this method, consider a typical pharmacological research question. Researchers are investigating whether four distinct drugs (Drug 1, Drug 2, Drug 3, and Drug 4) elicit different reaction times in human subjects. To ensure maximum control and statistical power, they employ a crossover design where five patients are tested under the influence of all four drugs, with sufficient washout periods between conditions to prevent carryover effects. Reaction time is the dependent variable measured in milliseconds, while the drug type is the within-subject factor.

The experimental structure dictates that each of the five patients provides four separate measurements--one under each drug condition. Therefore, the data structure is inherently dependent across the treatment variable. If we simply averaged the results for each drug and ran a standard ANOVA, we would overlook the powerful effect of controlling for baseline differences among patients. The repeated measures ANOVA is the only appropriate tool here to rigorously determine if the mean reaction time differs significantly across the four drugs, effectively controlling for the variability introduced by the specific patient.

Our primary goal is to use the statistical framework to test the central hypothesis of the study. Since each patient is measured on each of the four drug conditions, we must utilize the repeated measures ANOVA to ascertain if the true population mean reaction time differs across these treatments. The following steps outline the exact procedure for performing this analysis within the Python environment, transforming raw experimental data into interpretable statistical evidence.

Step 1: Data Structuring and Entry in Python

The initial and most critical step involves structuring the data correctly within a suitable Python data structure. For statistical analysis, the data must be in a 'long' format, meaning each row represents a single observation (a patient's response to a single drug), rather than a 'wide' format where each row represents a patient and columns represent the different drug treatments. We rely on the NumPy and Pandas libraries to efficiently create and manage this required data structure. The Pandas DataFrame is the standard container for this type of statistical data manipulation.

In the code below, we define three key variables: the subject identifier (`patient`), the within-subject factor level (`drug`), and the dependent measure (`response`). The use of NumPy functions like

`repeat` and `tile` ensures that we correctly map the patient IDs to all conditions and cycle through the drug levels sequentially, creating the necessary structure for the repeated measures analysis. This setup ensures that the statistical model can correctly identify which responses belong to which subject and which treatment level.

We import the necessary libraries and then instantiate the DataFrame. Observing the first ten rows confirms that the data is organized correctly, with Patient 1 having responses for Drugs 1, 2, 3, and 4, followed by Patient 2 for Drugs 1, 2, 3, and 4, and so on. This long format is absolutely essential for the `AnovaRM` function in `statsmodels` to correctly parse the within-subject dependencies inherent in the data structure.

import numpy as np
import pandas as pd

#create data
df = pd.DataFrame({'patient': np.repeat(, 4),
'drug': np.tile(, 5),
'response': })

#view first ten rows of data
df.head

patient drug response
0 1 1 30
1 1 2 28
2 1 3 16
3 1 4 34
4 2 1 14
5 2 2 18
6 2 3 10
7 2 4 22
8 3 1 24
9 3 2 20

Step 2: Execution of the Repeated Measures ANOVA using Statsmodels

Once the data is correctly prepared, the next step involves calling the appropriate statistical function. We leverage the power of the statsmodels library, specifically the AnovaRM class, which is designed for Repeated Measures ANOVA calculations. This function requires explicit declaration of the data source, the dependent variable, the subject identifier, and the within-subject factor(s).

Correctly specifying these parameters is paramount to achieving a valid statistical test.

In our execution, we pass the DataFrame `df` to the `data` parameter. The dependent variable, which is the outcome we are measuring, is specified as `'response'` (`depvar`). The crucial element that defines the repeated measures structure is the `subject` parameter, which is set to `'patient'`, instructing the function to treat measurements with the same patient ID as related. Finally, the within-subject factor, or the independent variable being manipulated, is specified as a list `` using the `within` parameter. The `.fit()` method executes the model estimation, calculating the necessary sums of squares and degrees of freedom required for the F-statistic.

The output generated by the `AnovaRM` function provides a concise table summarizing the results of the ANOVA. This output includes the calculated F-statistic, the degrees of freedom for the numerator (Num DF) and the denominator (Den DF), and the crucial P-value (Pr > F). These values collectively allow us to assess the statistical significance of the differences observed across the drug treatments, informing our decision regarding the null hypothesis. The simplicity of the code belies the complexity of the underlying calculations, which correctly account for the covariance structure.

from statsmodels.stats.anova import AnovaRM

#perform the repeated measures ANOVA
print(AnovaRM(data=df, depvar='response', subject='patient', within=).fit())

Anova
==========
F Value Num DF Den DF Pr > F
----
drug 24.7589 3.0000 12.0000 0.0000
==========

Step 3: Interpreting the ANOVA Results and Hypothesis Testing

The interpretation phase requires a clear understanding of the hypotheses tested by the ANOVA. The analysis operates under two competing statistical hypotheses, which dictate whether the observed differences are likely due to chance or a genuine treatment effect. Defining these hypotheses formally is the foundation of inferential statistics and allows us to contextualize the resulting F-test statistic and P-value.

The formal statements of the hypotheses for our experiment are:

The Null Hypothesis (H0): μ1 = μ2 = μ3 = μ4. This posits that the population mean reaction times

are equal across all four drug conditions. Statistically, it implies that the drug type has no effect on the average reaction time.

The Alternative Hypothesis (Ha): At least one population mean is different from the rest. This suggests that the drug type does significantly influence reaction time, meaning at least one drug has a different effect than the others.

In our specific output, we focus on the row corresponding to the 'drug' factor. The calculated F Value is **24.7589**, and the corresponding Pr > F (P-value) is **0.0000** (which conventionally signifies P < 0.001). The F-statistic is a ratio of the variance explained by the drug factor (treatment effect) to the unexplained variance (error). A large F-statistic suggests that the differences among the drug means are substantial compared to the variability within the conditions. Since the P-value (0.0000) is far smaller than the conventional significance level of 0.05 (or even 0.01), we confidently reject the Null Hypothesis. This leads to the conclusion that there are statistically significant differences in mean reaction times attributable to the type of drug administered.

Advanced Considerations: Degrees of Freedom and Effect Size

Beyond the F-statistic and P-value, the output provides essential information regarding the degrees of freedom (DF). The numerator DF (Num DF = 3.0000) reflects the number of treatment levels minus one (4 drugs - 1 = 3). The denominator DF (Den DF = 12.0000) represents the error degrees of freedom, which is calculated as the product of the numerator DF and the number of subjects minus one (3 * (5 patients - 1) = 12). These degrees of freedom are crucial for referencing the F-distribution table and for correctly reporting the results in academic literature.

While the `AnovaRM` output provides the core test statistics, a complete analysis often requires the calculation of an effect size measure, such as Partial Eta Squared (eta_p^2). Although not directly provided in this specific output table, calculating effect size is necessary to gauge the practical significance of the findings, moving beyond mere statistical significance. A large effect size would indicate that the drug factor accounts for a substantial proportion of the total variance in reaction time. In future versions of the analysis, researchers might incorporate post-hoc tests to determine precisely which pairs of drugs differ significantly, as the overall ANOVA only tells us that *at least one* difference exists.

Understanding the power of this test is enhanced by recognizing the source of the denominator in the F-ratio. Because the repeated measures design accounts for subject variability, the error term (Den DF) is smaller than it would be in an independent samples ANOVA. This reduction in the error variance is the reason why repeated measures designs often possess greater statistical power to detect smaller treatment effects, provided the correlation among the repeated measurements is substantial. The small P-value in this example is a testament to the effectiveness of the within-subjects approach.

Step 4: Formal Reporting of Results

The final step in the analytical process is to formally report the statistical findings in a clear, concise, and standardized manner, typically following APA guidelines for academic reporting. This ensures that the results are easily understood, verifiable, and integrated into the existing body of scientific knowledge. Reporting includes stating the type of test conducted, the sample size, the significant findings, and the critical test statistics (F-statistic, degrees of freedom, and P-value).

The standard format requires including the degrees of freedom in parentheses directly following the F statistic, followed by the calculated F-value, and then the exact or bounded P-value. Since our P-value is 0.0000, we report it as $p < 0.001$. This level of detail confirms the rigor of the analysis and allows peers to assess the validity of the conclusion drawn from the data. The narrative should link the statistical result back to the original research question concerning the effect of the four drugs on reaction time.

Here is an example of how the results of this repeated measures ANOVA should be formally reported in a research document:

A one-way repeated measures ANOVA was conducted on 5 individuals to examine the effect that four different pharmacological agents (drugs) had on measured response time. The analysis was conducted using the statsmodels library in Python, utilizing the long-format data structure appropriate for within-subjects designs.

The results demonstrated that the type of drug administered led to statistically significant differences in mean response time across the four conditions. The calculated F-statistic was highly significant ($F(3, 12) = 24.75887$, $p < 0.001$). This finding warrants further investigation using post-hoc procedures to determine the specific pairs of drug conditions that exhibit significant differences in efficacy, providing a foundation for determining which pharmacological agent performs optimally based on the outcome measure.

Conclusion and Further Analysis

The successful execution and interpretation of the repeated measures ANOVA in Python provides compelling evidence that the four drugs exert differential effects on patient reaction times. The utilization of the `AnovaRM` function within `statsmodels` is a straightforward yet powerful method for analyzing complex within-subjects experimental designs, provided the data is correctly structured in the long format. Researchers should always confirm that the assumptions underlying the repeated measures ANOVA, particularly sphericity, are met or addressed through appropriate correction methods in more complex or multivariate analyses.

For those seeking to expand their understanding, the subsequent logical step after finding a

significant overall F-test is to perform targeted post-hoc comparisons. These tests, such as Bonferroni or Tukey's HSD (Honest Significant Difference), are essential for controlling the Familywise Error Rate and identifying precisely which pair-wise differences between the four drug conditions are statistically significant. While the overall ANOVA confirms a difference exists, the post-hoc tests reveal the specific nature of that difference, completing the statistical picture.

The following tutorials and resources provide additional information on related statistical methodologies and advanced applications of repeated measures ANOVA: