How to Easily Perform Welch’s ANOVA in Python

Authored by

stats writer

December 6, 2025

The statistical procedure known as Welch's ANOVA is an advanced variant of the standard one-way Analysis of Variance (ANOVA) test. It is specifically employed when comparing the means of two or more independent groups, particularly in research settings where the assumption of equal population variances (homoscedasticity) is violated. Standard ANOVA yields unreliable results under conditions of heterogeneity of variances, making Welch's correction essential for maintaining statistical validity.

To execute a proper Welch's ANOVA in Python, a methodical approach involving data preparation, assumption testing, and specialized library usage is required. This process typically involves structuring data, importing robust statistical packages like Pingouin, computing the appropriate test statistic with adjusted degrees of freedom, and subsequently interpreting the test statistic and p-value to draw accurate conclusions about whether the groups exhibit statistically different means.

Welch's ANOVA is a vital alternative to the standard F-test used in traditional ANOVA when the foundational assumption of homogeneity of variances is violated in the data. By adjusting the degrees of freedom calculation, it provides a robust test for mean differences even when group variances are unequal.

The following detailed, step-by-step example demonstrates how to implement and interpret Welch's ANOVA effectively within the Python environment.

Introduction to Welch's ANOVA and Its Necessity

The Analysis of Variance (ANOVA) is a fundamental statistical technique used to test for differences between the means of two or more independent groups. Typically, standard one-way ANOVA relies on several crucial assumptions regarding the data distribution, one of the most critical being the homogeneity of variances. This assumption, often referred to as homoscedasticity, mandates that the variance within each group being compared must be approximately equal.

However, real-world data frequently violates this assumption. When group variances are significantly different--a condition known as heteroscedasticity--the results derived from standard ANOVA become unreliable, potentially leading to inaccurate statistical inferences regarding the null hypothesis. In such scenarios, the classic F-test is no longer valid, and an alternative approach is required to maintain the integrity of the statistical analysis.

This is where Welch's ANOVA, developed by Bernard Lewis Welch, serves as a powerful and robust alternative. Welch's ANOVA is specifically designed for situations where the assumption of equal variances across groups is violated. It achieves this robustness by adjusting the degrees of freedom used in the calculation of the F-statistic, effectively compensating for the heterogeneity.

This guide provides a detailed, step-by-step methodology for executing Welch's ANOVA using the Python programming language, ensuring valid results even when traditional ANOVA cannot be applied.

Setting Up the Experiment and Creating Sample Data

To illustrate the application of Welch's ANOVA, we will construct a hypothetical scenario involving three different pedagogical interventions. A research professor is investigating whether three distinct studying techniques--labeled Technique A, Technique B, and Technique C--have a differential effect on student performance as measured by exam scores. The goal is to determine if the population means of the exam scores across these three groups are statistically distinct.

In this controlled experiment, the professor randomly assigns 30 students, ensuring 10 participants are allocated to use each technique (Technique A, B, or C) for one week. This random assignment is crucial for maintaining internal validity and minimizing potential confounding variables. Following the intervention period, all students complete an identical exam designed to be of equal difficulty. The resulting dataset comprises 30 individual scores, categorized by the studying technique used, allowing for a comparative statistical analysis.

The raw exam scores collected from the 30 participants are structured into three lists, corresponding to the respective techniques. Analyzing this structure in Python first requires defining these lists, which will later be aggregated into a standard DataFrame for specialized statistical testing. The initial data representation is shown below:

A =
B =
C =

Preliminary Analysis: Understanding the Need for Welch's Test

Before proceeding with any form of ANOVA, researchers must evaluate whether the fundamental assumptions underlying the classical F-test are met. When comparing three or more groups, the primary concern, particularly regarding robustness, is the assumption of homogeneity of variances. If the spread of scores (variance) is substantially different between Technique A, Technique B, and Technique C, the statistical power and Type I error rate of the standard ANOVA are compromised.

A superficial examination of the raw scores often hints at variance inequality. For instance, Technique A shows a score range from 64 to 98, indicating a wide dispersion and likely high variance. Technique B, conversely, ranges from 82 to 97, suggesting a more concentrated group performance and potentially lower variance. These preliminary observations necessitate a formal

statistical test to definitively assess the equality of variances across the experimental groups.

Failing to test for this assumption and blindly applying standard ANOVA when variances are unequal dramatically increases the risk of drawing false conclusions, especially regarding the rejection of the null hypothesis (that all group means are equal). Therefore, the critical next step in our statistical procedure is the formal evaluation of the equal variance assumption using a dedicated statistical test.

Testing the Assumption of Homogeneity of Variances (Bartlett's Test)

To formally assess whether the variances of the three technique groups are equal, we employ Bartlett's test. Bartlett's test is sensitive to departures from normality but is highly effective for testing variance equality when the data distribution is known to be normal. The null hypothesis (H_0) for Bartlett's test states that all population variances are equal ($sigma_A^2 = sigma_B^2 = sigma_C^2$).

If the p-value resulting from the test statistic is less than a predetermined significance level (like $alpha = 0.05$), then we must reject the null hypothesis. Rejecting H_0 means there is sufficient statistical evidence to conclude that not all group variances are the same, thereby violating the core assumption of standard ANOVA. We utilize the readily available statistical functions within the Python scientific computing ecosystem, specifically the `scipy.stats` library, to perform this verification.

We can use the following code to perform Bartlett's test in Python, feeding it the three arrays of scores:

import scipy.stats as stats

#perform Bartlett's test
stats.bartlett(A, B, C)

BartlettResult(statistic=9.039674395, pvalue=0.010890796567)

The p-value (**0.01089**) derived from Bartlett's test is clearly smaller than our chosen significance level of $alpha = 0.05$. This result compels us to reject the null hypothesis that each group has the same variance. Consequently, the assumption of equal variances is violated, making it statistically necessary to bypass the traditional ANOVA and proceed directly to perform Welch's ANOVA.

Installing Required Libraries and Preparing Data for Analysis

The standard statistical packages in Python, such as NumPy and Pandas, provide robust tools for

data manipulation but do not natively offer a streamlined function for Welch's ANOVA. For specialized statistical tests like this, we rely on dedicated third-party libraries. The Pingouin library is an excellent choice, as it is built on top of NumPy and Pandas and provides a high-level API for various statistical routines, including the specific `welch_anova()` function we require.

Before proceeding, the Pingouin package must be installed within your Python environment. This is achieved using the standard package installer, `pip`:

pip install Pingouin

Once installed, the next crucial step is reshaping the data from its initial three-list format into a structure suitable for Pingouin and Pandas operations. Pingouin functions typically expect a long-format DataFrame, where one column represents the dependent variable (the scores) and another column specifies the independent variable or grouping factor (the technique). We achieve this transformation using the powerful capabilities of the Pandas library combined with NumPy for efficient array manipulation, creating a comprehensive DataFrame ready for analysis.

Executing Welch's ANOVA in Python using Pingouin

With the required packages imported and the data structured appropriately in a Pandas DataFrame, we can now execute the primary statistical test. The code below first imports the necessary libraries and constructs the DataFrame. Notice how `numpy.repeat` is used efficiently to assign the group labels ('a', 'b', 'c') to the corresponding exam scores, resulting in a dataset ready for analysis.

The core analysis is performed using the **welch_anova()** function from the Pingouin package. This function requires specifying the dependent variable (`dv='score'`), the between-subjects grouping factor (`between='group'`), and the data source (`data=df`). The output generated by Pingouin provides a summary table that includes the F-statistic, the adjusted degrees of freedom (ddof1 and ddof2), and the uncorrected p-value (p_{unc}).

import pingouin as pg
import pandas as pd
import numpy as np

#create DataFrame
df = pd.DataFrame({'score': ,
'group': np.repeat(, repeats=10)})

#perform Welch's ANOVA
pg.welch_anova(dv='score', between='group', data=df)

Source ddof1 ddof2 F p-unc np2
0 group 2 16.651295 9.717185 0.001598 0.399286

The overall p-value (**0.001598**) from the ANOVA table is less than the critical threshold of $alpha = 0.05$. This highly significant result leads us to reject the null hypothesis, confirming that the mean exam scores are not equal across the three studying techniques. The fractional value for ddof2 (16.651295) highlights the variance correction applied by the Welch method.

Interpreting Results and Performing Post-Hoc Analysis (Games-Howell)

A significant result from Welch's ANOVA indicates only that a difference exists somewhere among the groups. To determine exactly which pairs of means are significantly different, a post-hoc analysis is required. Given that the assumption of equal variances was violated, we must employ a post-hoc test that is robust to unequal variances, such as the Games-Howell post-hoc test.

The Games-Howell post-hoc test is performed using the `pairwise_gameshowell()` function from Pingouin. This test controls the family-wise error rate while allowing for heterogeneous variances and unequal sample sizes, making it the most appropriate choice following a significant Welch's ANOVA.

pg.pairwise_gameshowell(dv='score', between='group', data=df)

A B mean(A) mean(B) diff se T df pval
0 a b 77.3 91.8 -14.5 3.843754 -3.772354 11.6767 0.0072
1 a c 77.3 84.7 -7.4 3.952777 -1.872102 12.7528 0.1864
2 b c 91.8 84.7 7.1 2.179959 3.256942 17.4419 0.0119

By analyzing the outputted p-values (pval), we can interpret the specific mean differences:

The comparison between groups **a** and **b** yields a p-value of 0.0072, indicating a statistically significant difference.
The comparison between groups **a** and **c** yields a p-value of 0.1864, which is not statistically significant ($alpha > 0.05$).
The comparison between groups **b** and **c** yields a p-value of 0.0119, indicating a statistically significant difference.

In summary, the statistical evidence confirms that Technique B resulted in significantly different exam scores compared to both Technique A and Technique C. The means of Technique A and Technique C were not found to be significantly different from each other. The Games-Howell post-hoc test successfully identified the superior performance associated with Technique B while

properly accounting for the initial variance heterogeneity.