A Guide to Welch’s ANOVA in Python: Comparing Group Means with Unequal Variances

Authored by

Mohammed looti

November 5, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *A Guide to Welch’s ANOVA in Python: Comparing Group Means with Unequal Variances*. PSYCHOLOGICAL STATISTICS. Retrieved from https://statistics.arabpsychology.com/?p=10276

The Analysis of Variance (ANOVA) stands as a cornerstone in parametric statistics, primarily utilized to determine if there are significant differences between the means of three or more independent groups. It is a highly efficient method for comparing multi-group experimental outcomes. However, the reliability of the standard one-way ANOVA hinges entirely upon several strict assumptions regarding the data distribution, the most critical of which is the assumption of homogeneity of variances (homoscedasticity). When the spread of scores (variance) across the populations being compared is unequal, the standard F-test becomes inflated or deflated, rendering the resulting p-value potentially inaccurate and leading to unreliable statistical inference.

This challenge necessitates the use of a more robust alternative: Welch's ANOVA. This powerful variant is specifically engineered to handle situations where the population variances are unequal, a condition known as heteroscedasticity. Welch's method applies an advanced correction factor to the degrees of freedom calculation, making it far more trustworthy than the classic ANOVA F-test in non-ideal real-world data scenarios. This comprehensive guide details the process of executing a complete Welch's ANOVA procedure, from assumption testing to post-hoc analysis, using the powerful statistical capabilities available within Python. We will leverage specialized libraries to ensure both efficiency and statistical rigor.

The Necessity of Robust Statistical Testing

Parametric tests, including standard ANOVA, are built on idealized conditions. Violations of these core assumptions do not merely introduce minor errors; they can fundamentally shift the conclusions drawn from the data. The assumption of homogeneity of variances requires that the variability within each group being compared is roughly the same across all groups. This means that if we are comparing three different teaching methods, the spread of scores for Method A should be statistically similar to the spread of scores for Method B and Method C.

When variances differ significantly--when we observe heteroscedasticity--the pooling of variance estimates, which is central to the traditional ANOVA F-test calculation, becomes biased. If the sample sizes across groups are also unequal (though not the case in our example, it is common), this bias is severely compounded. Therefore, when the data clearly exhibits unequal variances, switching to Welch's ANOVA is not optional; it is a critical step for maintaining statistical integrity. Welch's test addresses this issue by using a separate estimate of variance for each group, thereby adjusting the error term and the degrees of freedom to provide a much more accurate test statistic.

The decision to use Welch's method must always be supported by a formal test of the homogeneity assumption, such as the Bartlett's test or Levene's test. These preliminary tests serve as gatekeepers, confirming whether the data meets the prerequisites for standard ANOVA or if the analysis must proceed with a non-standard, robust approach.

Defining the Research Context and Initial Data Setup

To demonstrate the practical application of Welch's ANOVA, we establish a clear research scenario. Our goal is to assess the impact of three distinct studying techniques (Techniques A, B, and C) on student performance, measured by a standardized exam score. This design involves one categorical independent variable (Studying Technique, with three levels) and one continuous dependent variable (Exam Score), making it a perfect candidate for one-way ANOVA procedures.

For the experimental setup, a professor enrolls 30 students and employs a randomized controlled design, assigning 10 students to utilize each specific technique for a defined period of study. The randomization ensures that the groups are independent, minimizing the risk of selection bias. Following the intervention phase, all 30 students take an exam of identical difficulty, generating the raw data points that will form the basis of our statistical comparison. The resulting exam scores are initially collected into separate lists, categorized by the assigned technique.

These lists represent the raw, ungrouped scores for each intervention technique. Note that even though the sample sizes (n=10 for each group) are equal, the variances within these lists are likely to differ, which is precisely what we must test before proceeding with the main analysis.

A =
B =
C =

These three lists--A, B, and C--contain the dependent variable data organized by the independent grouping factor. The next critical step involves using these lists to formally test the assumption of equal variances, which dictates whether we proceed with standard ANOVA or the more robust Welch's alternative.

Step 1: Validating Assumptions via Bartlett's Test (Execution)

Before executing any mean comparison test, we must statistically confirm the presence or absence of homogeneity of variances. We utilize the Bartlett's test, which is highly sensitive to non-normality but effective here, to formally assess this condition. The test evaluates the null hypothesis (H?) that all population variances are equal. If we find sufficient evidence to reject H?, we conclude that the variances are unequal, requiring the use of Welch's correction.

For statistical hypothesis testing, we typically set a significance level (α) at 0.05. If the resulting p-value is less than 0.05, we reject the null hypothesis of equal variances. Conversely, a p-value greater than 0.05 suggests that we fail to reject the null hypothesis, meaning the assumption holds and standard ANOVA could be used. We employ the widely used **scipy.stats** library in Python, passing the three raw data lists directly into the function:

```
import scipy.stats as stats

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

BartlettResult(statistic=9.039674395, pvalue=0.010890796567)
```

The resulting p-value from the Bartlett's test is approximately **0.0109**. Since this value is considerably smaller than our predetermined significance level of α **= 0.05**, we must formally reject the null hypothesis of equal variances. This rejection confirms that the assumption of homogeneity of variances has been violated by our dataset. Consequently, proceeding with the standard one-way ANOVA would yield unreliable results. We are thus statistically required to utilize Welch's ANOVA for the comparison of group means.

Step 2: Preparing Tidy Data for Pingouin

While the **scipy.stats** library handles assumption tests well using list inputs, advanced statistical packages often require data to be formatted in a "tidy" or "long" structure. To efficiently calculate **Welch's ANOVA**, we will employ the **Pingouin** package, a powerful statistical library built on top of NumPy and Pandas, specifically designed for user-friendly psychological and neuroscientific statistics. Unlike SciPy, **Pingouin**'s primary function for ANOVA requires the data to be structured as a single Pandas DataFrame where one column holds all the dependent variable scores, and another column identifies the group membership for each score.

Before we can import and utilize the package, ensure that **Pingouin** is installed within your Python environment. This is a crucial prerequisite for the subsequent steps, as it contains the necessary functions for both the Welch test and the appropriate post-hoc analysis.

```
pip install Pingouin
```

The next step involves transforming our raw lists (A, B, C) into this required long-format structure. This transformation is best handled using the **Pandas** library for data manipulation and **NumPy** for array operations, specifically to generate the repeating group labels. We concatenate all scores into a single column named **score** and create a corresponding column named **group** that labels which technique each score belongs to (A, B, or C).

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

#create DataFrame
```

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

# DataFrame 'df' is now ready for Welch's ANOVA
```

The resulting Pandas DataFrame, named **df**, is now structured appropriately for the **Pingouin** library. It contains the independent variable (**group**) and the dependent variable (**score**), setting the stage for the primary statistical test in the next phase.

Step 3: Calculating and Interpreting Welch's ANOVA F-Test

With the data prepared in the required long format, we proceed to execute the core analysis using **Pingouin**'s **welch_anova()** function. This function automatically implements the necessary Satterthwaite-type correction for the degrees of freedom, which is the mathematical mechanism that makes Welch's test robust against heteroscedasticity. Unlike standard ANOVA, which assumes integer degrees of freedom, Welch's method often produces fractional (non-integer) degrees of freedom, as seen in the output.

We must explicitly define the column names corresponding to the dependent variable (**dv='score'**) and the between-subjects factor (**between='group'**), linking them to our newly created DataFrame, **df**.

```
#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 output table summarizes the results of the Welch's ANOVA. The critical value for determining the significance of the overall test is the p-value, reported here under the column **p-unc** (unadjusted p-value). Our calculated p-value is **0.001598**. By comparing this value against the standard significance threshold of α **= 0.05**, we observe that the p-value is substantially lower than the threshold.

This low p-value leads us to reject the global null hypothesis, which states that all group means are equal (**H?:** μ**A =** μ**B =** μ**C**). The statistical conclusion is clear: there is a statistically significant difference in exam scores attributable to the three different studying techniques. The F-statistic of 9.717185, accompanied by the partial eta-squared (**np2**) of 0.399, suggests a strong effect size for the grouping variable. However, rejecting the global null hypothesis only tells us that *at least one* mean differs from another; it does not specify which pairs are significantly different. For that, we

require a post-hoc analysis.

Step 4: Pinpointing Differences with the Games-Howell Post-Hoc Test

Since the overall ANOVA test was significant, the final step involves performing pairwise comparisons between all possible group combinations (A vs. B, A vs. C, and B vs. C). This is known as post-hoc analysis. Crucially, because our initial assumption testing revealed unequal variances, we cannot use traditional post-hoc tests like Tukey's Honestly Significant Difference (HSD). Using Tukey's test would reintroduce the bias that Welch's test was designed to eliminate.

The statistically appropriate post-hoc procedure when variances are unequal is the Games-Howell post-hoc test. The Games-Howell test is robust to both unequal variances and unequal sample sizes, providing accurate confidence intervals and p-values for all pairwise comparisons. We utilize the **pairwise_gameshowell()** function from the **Pingouin** package, using the same DataFrame structure established in Step 2:

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

We analyze the **pval** column of the Games-Howell post-hoc test output to determine which specific pairs exhibit a statistically significant difference ($p < 0.05$):

Groups a vs. b: The p-value is 0.0072. Since this is less than 0.05, we conclude that the difference in exam scores between Technique A and Technique B is statistically significant.
Groups a vs. c: The p-value is 0.1864. Since this is greater than 0.05, we conclude that the difference between Technique A and Technique C is not statistically significant.
Groups b vs. c: The p-value is 0.0119. Since this is less than 0.05, we conclude that the difference between Technique B and Technique C is statistically significant.

Based on both the overall Welch's ANOVA and the subsequent Games-Howell post-hoc analysis, we can definitively conclude that Technique B leads to significantly higher exam scores compared to both Technique A and Technique C. Technique B, with the highest mean score, is identified as the most effective method among the three tested interventions.

Additional Resources for Advanced Statistical Analysis

For those seeking to deepen their understanding of robust statistical methods and advanced data

manipulation in Python, the following resources provide excellent supplementary material:

Detailed explanations of the theoretical foundations and implementation differences between standard ANOVA and Welch's ANOVA methods.
Comprehensive documentation for the Pingouin statistical package, including tutorials on other robust tests and effect size calculations.
Guides on advanced data manipulation techniques using the Pandas DataFrame structure, essential for preparing complex datasets for statistical modeling.