

A Comprehensive Guide to Welch's t-test in Stata: Comparing Means with Unequal Variances

Authored by
Mohammed looti

November 8, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *A Comprehensive Guide to Welch's t-test in Stata: Comparing Means with Unequal Variances*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=13655>

The comparison of means between two distinct and independent groups is a cornerstone of statistical inference. Typically, researchers rely on the [independent two-sample t-test](#) (often called Student's t-test). However, this procedure relies on a critical assumption: **homogeneity of variance** (or homoscedasticity). This assumption mandates that the spread or variability of the outcome variable must be virtually identical across the two populations from which the samples are drawn.

When the assumption of equal [variance](#) is violated--a frequent reality in observational and experimental data--the standard Student's t-test becomes statistically unreliable. Using the standard test under conditions of heteroscedasticity can lead to inflated Type I errors, potentially resulting in the false rejection of the null hypothesis. To address this common challenge, statisticians developed [Welch's t-test](#) (sometimes referred to as the unequal variances t-test or the Satterthwaite approximation). This robust modification ensures accurate inference even when population variances differ significantly.

This comprehensive guide provides a practical, step-by-step methodology for executing and interpreting [Welch's t-test](#) within the powerful [Stata](#) statistical software environment. We will cover the entire analytical workflow, including data preparation, exploratory visualization to confirm unequal variances, command execution, and formal reporting of the results.

Understanding the Rationale for Choosing Welch's t-test

Parametric statistical tests, such as the standard Student's t-test, assume that the populations being analyzed share fundamental distributional properties, particularly regarding their dispersion. When comparing group means, the underlying variability is a crucial factor. If the variability (variance) is substantially different between groups, the standard t-test calculates a pooled standard deviation that is biased, especially if the sample sizes are unequal. This bias compromises the test's validity and can severely distort the resulting p-value.

[Welch's t-test](#) addresses this limitation by abandoning the requirement for equal population variances. Instead, it calculates the test statistic using a separate variance estimate for each group. This methodological shift dramatically enhances the test's robustness when dealing with **heteroscedasticity**. A further critical adjustment is made to the calculation of the degrees of freedom (df). Welch's t-test employs the complex Satterthwaite equation, which often yields non-integer degrees of freedom. This adjustment ensures that the critical value used for hypothesis testing is correctly calibrated to account for the disparate variability observed between the two samples.

While non-parametric alternatives, such as the Mann-Whitney U test, are available when distributional assumptions are violated, [Welch's t-test](#) is generally preferred if the data are approximately normally distributed but exhibit unequal variances. It retains the superior statistical

power associated with parametric methods while successfully correcting for the failure of the homogeneity of variance assumption. For many real-world data scenarios, Welch's test provides the most powerful and statistically sound approach for comparing group means.

Setting Up the Stata Environment and Loading Data

To demonstrate the proper application of Welch's t-test, we will use a relevant, public dataset. Our illustration utilizes the *fuel3* dataset, which is conveniently hosted in the Stata Press online repository. This dataset is designed to investigate the effect of a specific fuel treatment on the mean miles per gallon (mpg) of a fleet of automobiles. The data includes measurements for 24 cars: 12 cars assigned to the treatment group (coded as Group 1) and 12 cars designated as the control group (coded as Group 0).

The essential first step in any [Stata](#) analysis is securely loading the required data. Since the *fuel3* dataset is available online, Stata facilitates immediate remote access. To load this dataset directly into your current session, execute the following command in the Stata Command window:

use <http://www.stata-press.com/data/r13/fuel3>

After loading the data, it is crucial to perform an initial inspection to verify that the variables are correctly structured and interpreted. We can quickly examine the raw data structure, confirming all observations and variable types, by using the standard `list` command. This verification step ensures the presence and correct definition of our two core variables: the continuous outcome variable (**mpg**) and the binary grouping variable (**treated**).

list

The resulting output confirms that the dataset contains 24 observations, each successfully categorized by the `treated` status (0 for control, 1 for treatment) alongside the corresponding measured `mpg` value, preparing the dataset for the next analytical stage.

```
. use http://www.stata-press.com/data/r13/fuel13
```

```
. list
```

	mpg	treated
1.	20	0
2.	23	0
3.	21	0
4.	25	0
5.	18	0
6.	17	0
7.	18	0
8.	24	0
9.	20	0
10.	24	0
11.	23	0
12.	19	0
13.	24	1
14.	25	1
15.	21	1
16.	22	1
17.	23	1
18.	18	1
19.	17	1
20.	28	1
21.	24	1
22.	27	1
23.	21	1
24.	23	1

Visual Data Exploration: Confirming Unequal Variances

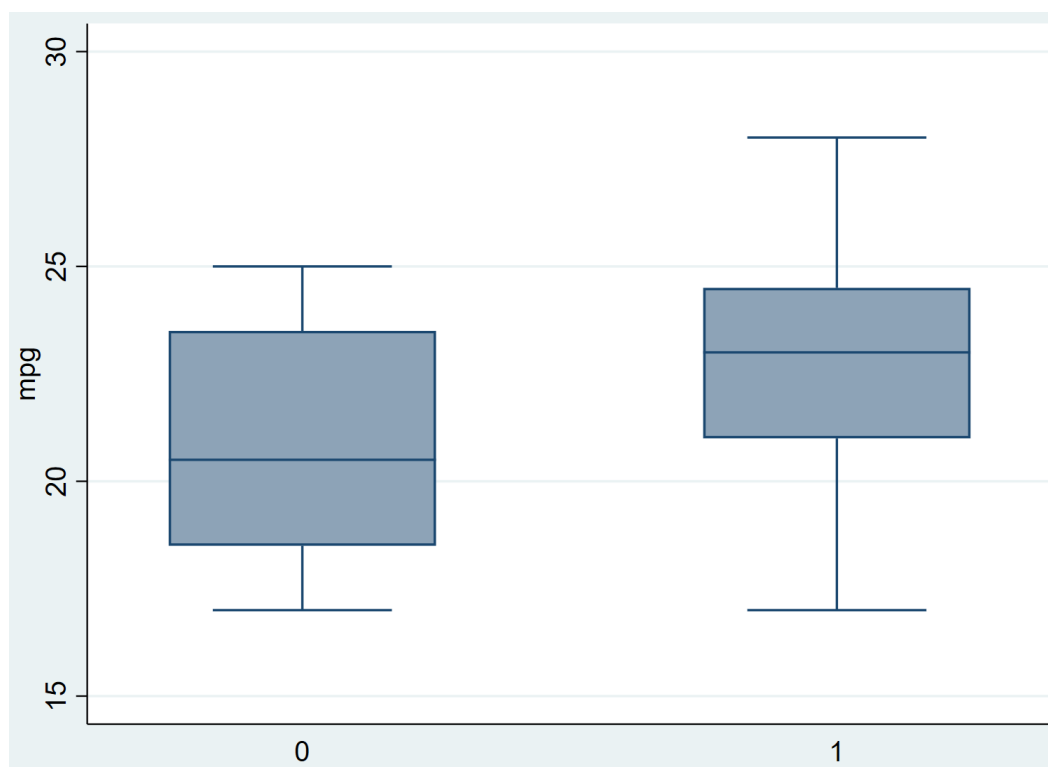
Before proceeding with formal inferential statistics, best analytical practices require a robust visual inspection of the data. Visualization offers an intuitive and preliminary assessment of the distribution of the outcome variable within each group. This step is essential not only for assessing differences in central tendency but, more critically for our purpose, for identifying potential differences in [variance](#).

The box plot is an exceptionally effective visualization tool for comparing distributions across groups. By generating side-by-side box plots for the miles per gallon (mpg) variable, stratified by the treated and untreated status, we gain immediate insight into their comparative spreads. We instruct Stata to create these comparative plots using the following command, specifying the

`over ( )` option to stratify the visualization by the grouping variable:

graph box mpg, over(treated)

The resulting graph displays key distributional summaries--the median, the interquartile range (IQR), and any potential outliers--for both groups. This visual comparison is the primary factor in determining whether the assumption of homoscedasticity is tenable, or whether we must proceed with the statistically sound [Welch's t-test](#).



Reviewing the box plots reveals two important features. First, the median line for Group 1 (treated) appears visibly higher than that for Group 0 (untreated), suggesting a numerical difference in mean mpg. More importantly, we observe a distinct difference in the vertical spread of the boxes. The box plot representing Group 1 is significantly narrower than the box plot for Group 0. This graphical observation indicates a smaller Interquartile Range for the treated cars, providing strong visual evidence that the variance in mpg for the treated group is less than the variance for the untreated group. This visual confirmation of unequal variability justifies the selection and use of Welch's t-test.

Executing the Welch's t-test Command in Stata

Once the data is prepared, visualized, and the necessity for a robust test has been established, executing [Welch's t-test](#) in Stata is a simple process. The fundamental command for conducting a

two-sample hypothesis test on means is `ttest`. To seamlessly invoke the Welch modification, we simply append the dedicated `welch` option to the standard syntax.

The required syntax for this procedure is structured clearly: the command is applied to the outcome variable, stratified by the grouping variable, with the specific modification explicitly requested via the option:

`ttest variable_to_measure, by(grouping_variable) welch`

Applying this structure to our practical example using the *fuel3* dataset, where we measure `mpg` across the categories defined by the `treated` variable, the specific command is:

`ttest mpg, by(treated) welch`

Upon execution, Stata performs the two-sample t-test without assuming equality of variances. Stata utilizes the separate variance estimates to compute the t-statistic and determines the appropriate degrees of freedom (df), which are typically non-integer, through the Satterthwaite approximation method. The resulting output table provides all the essential statistics required for formal hypothesis testing and scholarly reporting.

`. ttest mpg, by(treated) welch`

Two-sample t test with unequal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	12	21	.7881701	2.730301	19.26525	22.73475
1	12	22.75	.9384465	3.250874	20.68449	24.81551
combined	24	21.875	.6264476	3.068954	20.57909	23.17091
diff		-1.75	1.225518		-4.28369	.7836902

diff = mean(0) - mean(1)

t = -1.4280

Ho: diff = 0

Welch's degrees of freedom = 23.2465

Ha: diff < 0

Pr(T < t) = 0.0833

Ha: diff != 0

Pr(|T| > |t|) = 0.1666

Ha: diff > 0

Pr(T > t) = 0.9167

Interpreting the Stata Output for Robust Inference

The output generated by the `ttest` command with the `welch` option is comprehensive, detailing descriptive statistics, precise [confidence interval](#) estimates, and the inferential results of the hypothesis test. Effective interpretation necessitates a careful examination of the three main

sections presented in the table.

The top section summarizes the descriptive statistics for both groups. For Group 0 (untreated), the mean mpg was calculated as **21.00**. The associated 95% [confidence interval](#) for the true population mean of this group was (19.26525, 22.73745). Conversely, Group 1 (treated) showed a numerically higher mean mpg of **22.75**, with its 95% [confidence interval](#) ranging from (20.68449, 24.81551). This initial comparison confirms the numerical superiority of the treatment group, as initially suggested by the box plots.

The inferential core of the output is found in the bottom section, which focuses on the mean difference (calculated as Group 0 minus Group 1), resulting in a value of **-1.75**. The 95% [confidence interval](#) for this true difference in population means is reported as **(-4.28369, 0.7836902)**. A critical step in interpretation is assessing whether this interval encompasses zero. Since the interval spans both negative and positive values (i.e., it contains zero), this strongly suggests that the true difference between the population means could plausibly be zero, indicating a lack of statistical significance.

Finally, we examine the formal hypothesis test results. The test statistic, t , for [Welch's t-test](#) is **-1.4280**. Assuming a two-sided alternative hypothesis ($H_a: \text{diff} \neq 0$), we focus on the corresponding [p-value](#), which is **0.1666**. Using the conventional significance level (α) of 0.05, we compare the calculated [p-value](#) to this threshold. Since 0.1666 is substantially greater than 0.05, we must consequently fail to reject the null hypothesis. The analysis concludes that there is insufficient statistical evidence to assert that the fuel treatment caused a statistically significant difference in mean mpg between the two groups.

Reporting Statistical Results and Conclusions

The final stage of any rigorous data analysis involves formally reporting the findings of the statistical test in a standardized and transparent manner. A proper statistical report must include all relevant parameters necessary for reproducibility and clarity, specifically detailing the test utilized, the degrees of freedom, the calculated test statistic, and the final p-value.

Below is an example illustrating how the results derived from the *fuel3* dataset analysis should be formally reported, adhering to established statistical conventions:

A [Welch's t-test](#) was conducted to investigate whether a specialized fuel treatment resulted in a statistically significant difference in mean miles per gallon (mpg) compared to a control group. The robust Welch modification was selected due to visual evidence indicating unequal population variance between the two groups ($n = 12$ in each group).

The Welch's t-test revealed that the difference in means was **not** statistically significant ($t =$

-1.4280, $p = 0.1666$). The control group reported a mean $M = 21.00$, while the treatment group reported a mean $M = 22.75$.

The 95% [confidence interval](#) for the true difference in mean mpg (untreated minus treated) was calculated as $(-4.28369, 0.7836902)$. As this interval includes zero, it reinforces the conclusion that, at the 0.05 significance level, there is no statistically discernible difference in fuel efficiency attributable to the treatment.