

How to Perform Welch's t-test in Stata: A Step-by-Step Guide

Authored by
stats writer

March 8, 2026

RECOMMENDED CITATION

stats writer (2026). *How to Perform Welch's t-test in Stata: A Step-by-Step Guide*.

PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=134656>

Theoretical Foundations of Welch's t-test

In the realm of inferential statistics, **Welch's t-test**, also known as the unequal variances **t-test**, serves as a robust alternative to the standard independent samples t-test. This particular statistical procedure is specifically engineered to determine if there is a **statistically significant** difference between the **means** of two independent groups. Unlike its more traditional counterpart, Welch's method does not rely on the strict assumption that the two populations from which the samples are drawn possess equal **variance**. This flexibility makes it an indispensable tool for researchers dealing with real-world data, where **homoscedasticity** is frequently violated.

The mathematical brilliance of **Welch's t-test** lies in its calculation of the **degrees of freedom**. While the standard t-test uses a simple additive approach for degrees of freedom based on sample sizes, Welch's test employs the **Welch-Satterthwaite equation**. This formula adjusts the degrees of freedom based on both the sample sizes and the specific variances of each group. By doing so, the test provides a more accurate **p-value**, thereby controlling the **Type I error rate** more effectively than the Student's t-test when variances are unequal. This adjustment is crucial because using a standard t-test on heteroscedastic data can lead to unreliable results and misleading conclusions.

When working within **Stata**, implementing this test is remarkably straightforward, yet it requires a deep understanding of when and why it should be applied. The software handles the complex adjustments to the **degrees of freedom** automatically once the user specifies the appropriate option. It is generally recommended by modern statisticians to use Welch's t-test as a default when comparing two means, as it performs nearly identically to the Student's t-test when variances are equal but remains valid when they are not. Consequently, mastering this command in **Stata** is a fundamental skill for any data analyst or academic researcher.

Distinguishing Welch's t-test from the Classic Student's t-test

The primary distinction between **Welch's t-test** and the classic Student's **t-test** centers on the assumption of **homoscedasticity**. The Student's t-test operates under the premise that the variability within each group is approximately the same. When this assumption holds true, the pooled **variance** estimate used by the Student's t-test is highly efficient. However, in many biological, social, and economic studies, sample sizes may differ significantly between groups, and the underlying spread of data points can vary wildly. In such instances, the standard t-test becomes unreliable, often producing **confidence intervals** that are either too narrow or too wide, which compromises the integrity of the findings.

Researchers often utilize **Welch's t-test** because it handles the **Behrens-Fisher problem**, which refers to the difficulty in testing the hypothesis that two means are equal when the variances are

unknown and not necessarily equal. By calculating separate variances for each sample rather than pooling them, Welch's method provides a more localized and accurate representation of the data's dispersion. This ensures that the resulting test statistic, t , is a more faithful reflection of the actual difference between the means relative to the inherent noise in the data. For this reason, many statistical software packages, including **Stata**, have made the Welch option easily accessible for routine analysis.

It is also important to consider the normally distributed nature of the data. Both tests assume that the populations are normally distributed. While the t-test is somewhat robust to minor departures from normality--especially with larger sample sizes due to the Central Limit Theorem--it is always prudent to perform diagnostic checks on your data before proceeding. If your data is highly skewed or contains significant outliers, even Welch's test might not be the most appropriate choice, and you might consider non-parametric alternatives. However, for most continuous data sets where variance is the only major concern, Welch's remains the gold standard.

Initial Data Loading and Inspection Procedures in Stata

Before executing any statistical test in **Stata**, the first step is to correctly import and examine the dataset. In this tutorial, we will utilize a classic example dataset titled *fuel3*, which is hosted by Stata Press. This dataset is ideal for demonstrating Welch's t-test because it provides a clear grouping variable (treated vs. non-treated cars) and a continuous outcome variable (miles per gallon, or mpg). To begin, you must load the data directly from the online repository by entering the following command into the **Stata** command line:

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

Once the dataset is successfully loaded into your workspace, it is vital to perform a visual inspection of the raw values. This allows you to verify that the variables have been imported correctly and to understand the structure of the observations. The **list** command is the most direct way to view the records in the Results window. By reviewing the first few rows, you can confirm that the "mpg" variable contains numerical values and the "treated" variable is a binary indicator (typically 0 and 1). To see the data, type:

```
list
```

```
. 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

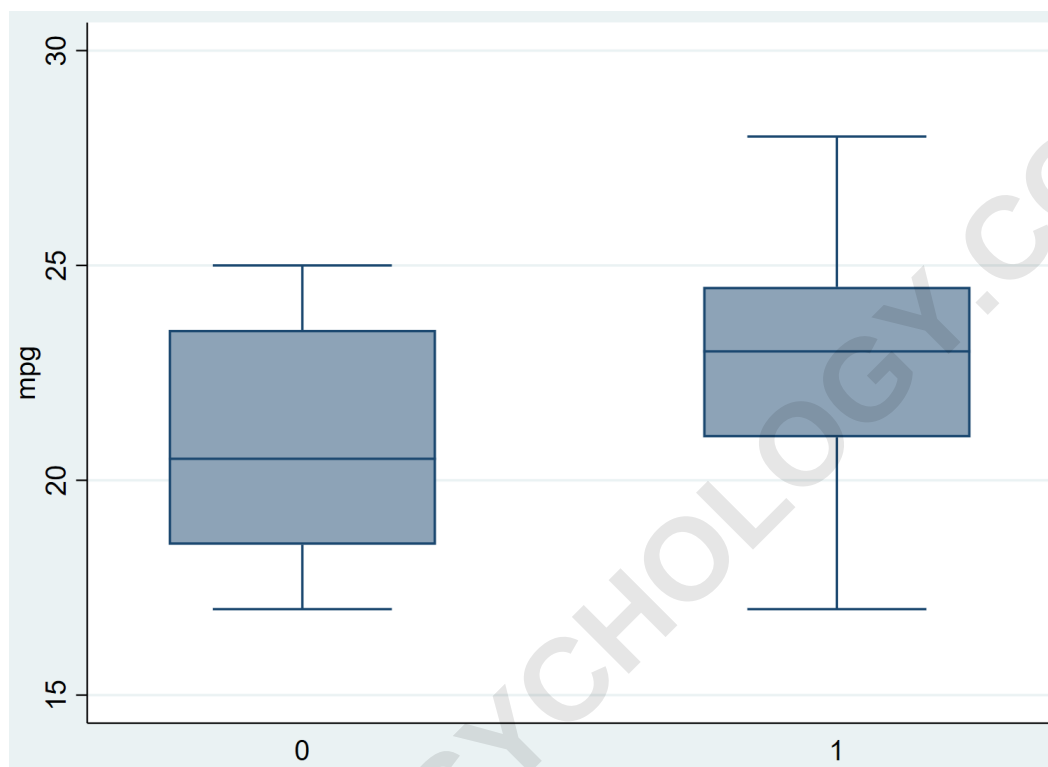
As shown in the output, the dataset consists of 24 observations in total, split evenly between two groups. Each group contains 12 cars, allowing us to compare the efficiency of those that received a specific fuel treatment against those that did not. Observing the data at this stage helps you identify any potential missing values or data entry errors that could skew your **means** or affect the **variance** calculations later in the analysis. Thorough data cleaning and inspection are the hallmarks of professional statistical work.

Exploratory Data Analysis and Visualization with Box Plots

A critical phase in any data analysis workflow is exploratory data analysis (EDA). Before calculating a **p-value**, you should always visualize the distribution of your variables. For comparing **means** between two groups, **box plots** (or box-and-whisker plots) are exceptionally useful. They

provide a concise summary of the median, quartiles, and potential outliers, while also offering a visual hint about whether the **variance** of the two groups is roughly equal. In **Stata**, you can generate side-by-side box plots using the following command:

```
graph box mpg, over(treated)
```



When interpreting these **box plots**, pay close attention to the "height" of the boxes and the length of the whiskers. In our specific example involving fuel treatments, you will likely notice that the mpg for Group 1 (the treated cars) appears generally higher than that of Group 0. Furthermore, the vertical spread of the box for Group 1 may appear narrower than that of Group 0, suggesting that the **variance** is not consistent across both groups. This visual confirmation is exactly why **Welch's t-test** is the preferred choice for this specific analysis.

Beyond looking at the **variance**, visualize the symmetry of the boxes to check if the data is **normally distributed**. While box plots do not provide a definitive test for normality, a median line that is centered within the box and whiskers of roughly equal length are indicators of a symmetric distribution. If the visualization reveals severe skewness or extreme outliers, you might need to transform your data or use a **non-parametric** test like the **Mann-Whitney U test**. However, given the visual evidence of unequal spreads here, we proceed with the Welch-adjusted approach.

Executing the Welch's t-test Syntax in the Stata Environment

Once you have visually inspected the data and determined that a comparison of means is appropriate despite unequal variances, you can proceed to the main analysis. **Stata** uses the **ttest** command for all variations of the t-test, and the specific syntax for **Welch's t-test** involves adding the "welch" option at the end of the command string. This tells Stata not to use the pooled variance estimate and instead calculate the Satterthwaite **degrees of freedom**. The general syntax structure is as follows: **ttest variable_to_measure, by(grouping_variable) welch**.

For our fuel treatment study, the "variable_to_measure" is **mpg** and the "grouping_variable" is **treated**. By running the command below, Stata will produce a comprehensive table containing the sample means, standard errors, and the final test results. It is important to remember that the "by" option is used when your data is in "long" format, meaning one column contains the measurement and another column identifies the group. To execute the test, enter:

```
ttest mpg, by(treated) welch
```

```
. 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

Ha: diff != 0

Ha: diff > 0

Pr(T < t) = 0.0833

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

Pr(T > t) = 0.9167

The resulting output is dense with information. At the top of the table, you will see a summary for each group, including the number of observations (N), the **mean**, the **standard error**, and the **standard deviation**. Below this, Stata provides a comparison of the means, the calculated t-statistic, and the adjusted **degrees of freedom**. Notice that the degrees of freedom in a Welch's test are often not an integer (e.g., 19.34 instead of 22), which is a clear indicator that the Welch adjustment has been applied correctly.

Detailed Interpretation of the Stata Statistical Output

Interpreting the **Welch's t-test** output requires looking at several key metrics simultaneously. First, examine the **means**: for Group 0 (non-treated), the average mpg was 21, while for Group 1 (treated), the average was 22.75. This suggests a raw difference of 1.75 mpg. However, the **standard deviation** for Group 0 is 2.73, which is considerably higher than the 1.60 observed in Group 1. This discrepancy in spread confirms that our initial decision to use the Welch option was statistically sound, as the variances are clearly not equal.

Next, focus on the bottom section of the output, which presents three different **p-values** based on different **alternative hypotheses**. The center p-value, labeled **Pr(|T| > |t|)**, is for a two-tailed test, which evaluates the **null hypothesis** that the means are equal. In our case, this value is 0.1666. Since this **p-value** is greater than the standard alpha level of 0.05, we fail to reject the **null hypothesis**. This means that, despite the observed difference in means, we do not have enough evidence to conclude that the fuel treatment significantly improves gas mileage.

The t-statistic of -1.4280 further supports this conclusion. A t-statistic tells us how many standard errors the difference in means is away from zero. In this instance, the difference is relatively small compared to the variability in the data. If the **p-value** had been below 0.05, we would have concluded that the result was **statistically significant**. However, the high degree of overlap between the two groups--exacerbated by the high variance in the untreated group--results in a non-significant finding.

Analyzing Confidence Intervals and Mean Differences

While **p-values** tell us whether an effect exists, **confidence intervals** provide essential information about the magnitude and precision of that effect. In the **Stata** output for **Welch's t-test**, look for the row labeled "combined" or the specific row representing the difference between the groups. The 95% **confidence interval** for the difference in means is reported as (-4.28369, 0.7836902). This interval represents the range of values that we are 95% confident contains the true difference between the population means.

A critical observation here is that the **confidence interval** includes the value of zero. In frequentist statistics, if the interval for the difference between two **means** spans zero, it is equivalent to saying that the **p-value** is greater than 0.05. This reinforces our earlier conclusion: because the interval goes from a negative value (-4.28) to a positive value (0.78), we cannot definitively say whether the treatment group performs better or worse than the control group in the broader population. The wide range of the interval also suggests a high degree of uncertainty, likely due to the small sample size (n=12 per group).

Understanding these intervals is vital for **effect size** estimation. Even if a result is not **statistically**

significant, the **confidence interval** tells us what the plausible range of the effect could be. For instance, the treatment could potentially increase mpg by as much as 4.28 or decrease it by 0.78. To narrow this interval and obtain a more precise estimate, a researcher would likely need to increase the sample size in a follow-up study. Reporting both the point estimate and the interval is considered best practice in modern scientific literature.

Professional Reporting and Contextualizing the Findings

The final step in any statistical analysis is the professional reporting of the results. When writing for a journal or a technical report, you must include the name of the test, the test statistic (t), the **degrees of freedom** (df), and the **p-value**. Because **Welch's t-test** uses a specific adjustment, it is common practice to report the decimal degrees of freedom provided by **Stata**. An example of a formal report would be: "A Welch's t-test was conducted to compare the fuel efficiency of treated and untreated cars. There was no **statistically significant** difference in the means of the treated group ($M = 22.75$, $SD = 1.60$) and the untreated group ($M = 21.00$, $SD = 2.73$); $t(17.8) = -1.43$, $p = 0.167$."

In addition to the basic statistics, providing context regarding the **confidence interval** is highly recommended. You might add: "The 95% **confidence interval** for the difference in means ranged from -4.28 to 0.78 mpg, further indicating that the effect of the fuel treatment was not distinguishable from zero at the 5% significance level." This level of detail allows readers to evaluate the robustness of your study and the practical implications of your findings. It also demonstrates a sophisticated understanding of how **variance** and sample size interact to produce statistical outcomes.

Finally, always conclude by discussing the limitations and the assumptions of the test. While **Welch's t-test** is robust to unequal variances, it still assumes that the data is **normally distributed**. If there were concerns about normality, you should mention that diagnostic plots were reviewed. In the case of our fuel study, the lack of significance may simply be a matter of low **statistical power**. By following these structured steps in **Stata**--from loading and visualization to execution and reporting--you ensure that your comparative analysis is accurate, transparent, and academically rigorous.