How to Easily Perform a T-Test with Unequal Sample Sizes

Authored by

stats writer

November 28, 2025

ARABPSYCHOLOGY.COM

RECOMMENDED CITATION

stats writer (2025). *How to Easily Perform a T-Test with Unequal Sample Sizes*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=100764

The t-test is a foundational tool in inferential statistics, used primarily to determine if there is a statistically significant difference between the means of two independent groups. Successfully performing this test requires careful consideration of the underlying assumptions, particularly those related to the structure and characteristics of the data collected. A common procedural concern for students and researchers is handling situations where the number of observations in each group--the sample sizes--are not equal.

Performing a t-test when the sample sizes are unequal is certainly possible and frequently necessary in real-world data analysis. The key steps involve rigorous calculations to ensure that the resulting t-statistic accurately reflects the distribution under the unequal conditions. This procedure generally requires calculating the degrees of freedom, determining the critical t-value based on the desired level of significance, calculating the pooled variance (if appropriate), and finally, calculating the empirical t-statistic. Only when the calculated t-statistic exceeds the critical t-value can we confidently state that the observed difference between the two sample populations is statistically significant.

Addressing the Core Question: T-Tests and Unequal Samples

A frequent query in statistical methodology centers on the rigidity of the t-test assumptions. Specifically, many analysts ask: **Is it permissible to conduct an independent samples t-test when the sample sizes of the two comparison groups are different?** The definitive statistical answer is **Yes**. The standard independent samples t-test, often referred to as the Student's t-test, does not inherently assume or require equal sample sizes ($n_1 = n_2$). The test is robust enough to handle imbalances, provided other, more critical assumptions are met.

While unequal sample sizes themselves do not invalidate the test, they can significantly amplify the negative effects of violating other core assumptions. The presence of imbalance, especially when severe (e.g., $n_1=500$ and $n_2=20$), means that the smaller group will have far less influence on the estimation of pooled parameters, potentially leading to increased susceptibility to outliers or skewness in the larger group. Therefore, while unequal sample sizes are tolerated, they necessitate a heightened vigilance regarding other data properties.

The true statistical complication arises not from sample size disparity, but from the issue of unequal population variances. The classical Student's t-test assumes that the population variances from which the two samples are drawn are equal--an assumption known as homogeneity of variances. If this assumption is violated, particularly when combined with unequal sample sizes, the calculated t-statistic and associated p-value become unreliable, potentially leading to incorrect conclusions regarding the null hypothesis.

ARABPSYCHOLOGY.COM

The Critical Role of Variance Assumptions

In statistics, the assumption of homogeneity of variances is paramount when using the standard independent samples t-test. When variances are equal, we can "pool" the sample variances to obtain a single, more stable estimate of the common population variance. This pooling procedure provides the foundation for calculating the test statistic and the degrees of freedom ($df = n_1 + n_2 - 2$). However, if the population variances are genuinely unequal--a state called heteroscedasticity--pooling the variances introduces bias into the denominator of the t-statistic formula, compromising the results.

When unequal variances are suspected or confirmed through tests like Levene's Test or the F-test, or simply when the sample sizes are drastically different, the statistical community recommends abandoning the standard Student's t-test in favor of an alternative procedure: the Welch's t-test. The Welch's t-test is specifically designed for situations where the assumption of equal variances is violated. It performs a separate variance estimate for each sample and adjusts the degrees of freedom using the Satterthwaite approximation, resulting in a more accurate and robust test statistic under heteroscedastic conditions.

The choice between the Student's t-test and the Welch's t-test is critical, especially when sample sizes are not balanced. When sample sizes are equal, the standard t-test is relatively robust against moderate violations of the variance assumption. However, when sample sizes are unequal, the standard t-test becomes highly sensitive to heteroscedasticity, making the use of the Welch's t-test the safer and preferred default option. The following detailed examples demonstrate the practical difference between these two tests when facing unequal sample sizes under varying variance conditions.

Case Study 1: Unequal Sample Sizes with Equal Variances (Homoscedasticity)

Let us consider a scenario where a researcher is comparing the effectiveness of two different educational programs designed to improve student performance on a standardized exam. The programs are tested on groups of drastically different sizes, but based on prior knowledge or preliminary testing, the researcher assumes that the inherent variability (variance) of scores across the populations is similar.

The results gathered from the two programs are summarized below, illustrating a large disparity in sample sizes ($n_1=500$ vs $n_2=20$) while maintaining identical standard deviations, thus supporting the assumption of equal population variances:

Program 1:

ARABPSYCHOLOGY.COM

n (sample size): 500
x? (sample mean): 80
s (sample standard deviation): 5

Program 2:

n (sample size): 20
x? (sample mean): 85
s (sample standard deviation): 5

The goal is to determine if the observed difference in sample means (80 vs. 85) is significant enough to reject the null hypothesis that the population means are equal. Since the population standard deviations are hypothesized to be equal (both equal to 5), both the independent samples t-test (assuming equal variances) and the Welch's t-test (not assuming equal variances) should yield similar results, illustrating the robustness of the standard t-test when the equal variance assumption holds true, even with unequal sample distribution.

Statistical Implementation in R (Equal Variances)

To visualize the score distributions and subsequently perform the hypothesis tests, we utilize the R statistical programming environment. The following code simulates data matching the characteristics described above and generates a boxplot for visual comparison. Note how the variability (spread) appears relatively consistent between the two groups, despite the vastly different sample counts.

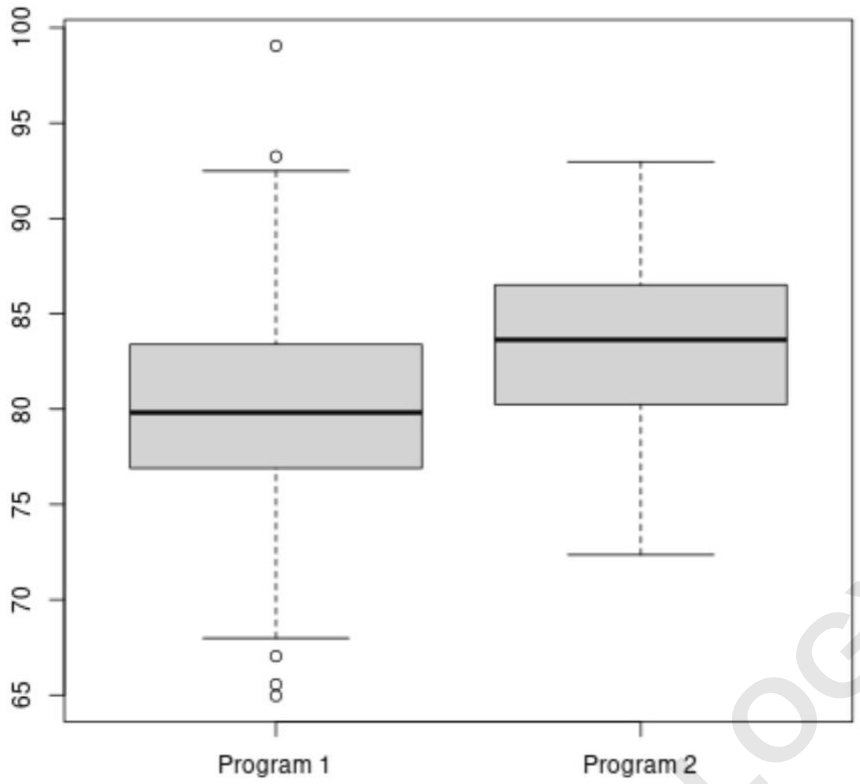
The following code shows how to create a boxplot in R to visualize the distribution of exam scores for each program:

```
#make this example reproducible
set.seed(1)

#create vectors to hold exam scores
program1 <- rnorm(500, mean=80, sd=5)
program2 <- rnorm(20, mean=85, sd=5)

#create boxplots to visualize distribution of exam scores
boxplot(program1, program2, names=c("Program 1","Program 2"))
```

ARABPSYCHOLOGY.COM



Next, we execute both the standard independent samples t-test (setting `var.equal=TRUE`) and the Welch's t-test (setting `var.equal=FALSE`) to compare their results under these ideal conditions:

#perform independent samples t-test
t.test(program1, program2, var.equal=TRUE)

Two Sample t-test

data: program1 and program2
t = -3.3348, df = 518, p-value = 0.0009148
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-6.111504 -1.580245
sample estimates:
mean of x mean of y
80.11322 83.95910

#perform Welch's two sample t-test
t.test(program1, program2, var.equal=FALSE)

Welch Two Sample t-test

ARABPSYCHOLOGY.COM

data: program1 and program2
t = -3.3735, df = 20.589, p-value = 0.00293
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-6.219551 -1.472199
sample estimates:
mean of x mean of y
80.11322 83.95910

The standard independent samples t-test yielded a p-value of **0.0009**, while the Welch's t-test returned a p-value of **0.0029**. Both values are significantly below the conventional alpha level of 0.05. Consequently, we would reject the null hypothesis in both cases, concluding that a statistically significant difference exists in the mean exam scores between the two programs. This example confirms that even with severely unequal sample sizes, when the assumption of equal variances holds true, both t-test methods lead to the same substantive conclusion.

Case Study 2: Unequal Sample Sizes and Unequal Variances (Heteroscedasticity)

Now, let us examine a scenario where the unequal sample sizes are paired with unequal population variances--the precise condition where the standard Student's t-test is most likely to fail. We maintain the same sample sizes ($n_1=500$ and $n_2=20$) and mean difference (80 vs. 85), but drastically increase the standard deviation for Program 1, introducing heteroscedasticity.

The results for this modified hypothetical experiment are:

Program 1:

n (sample size): 500
x? (sample mean): 80
s (sample standard deviation): 25 (High Variability)

Program 2:

n (sample size): 20
x? (sample mean): 85
s (sample standard deviation): 5 (Low Variability)

Here, the larger sample (Program 1, $n=500$) exhibits much greater variability than the smaller sample (Program 2, $n=20$). Because the larger sample has the higher variance, the standard t-test's pooled variance estimate will be heavily inflated by the variability of Program 1, making the

ARABPSYCHOLOGY.COM

test overly conservative and less likely to detect a true difference in means. This situation demonstrates the necessity of using the Welch's adjustment when both sample size imbalance and variance inequality coexist.

Statistical Implementation in R (Unequal Variances)

We once again use R to simulate the data and visualize the difference. Notice the dramatic change in the boxplot for Program 1, reflecting its much larger standard deviation (25) compared to Program 2 (5).

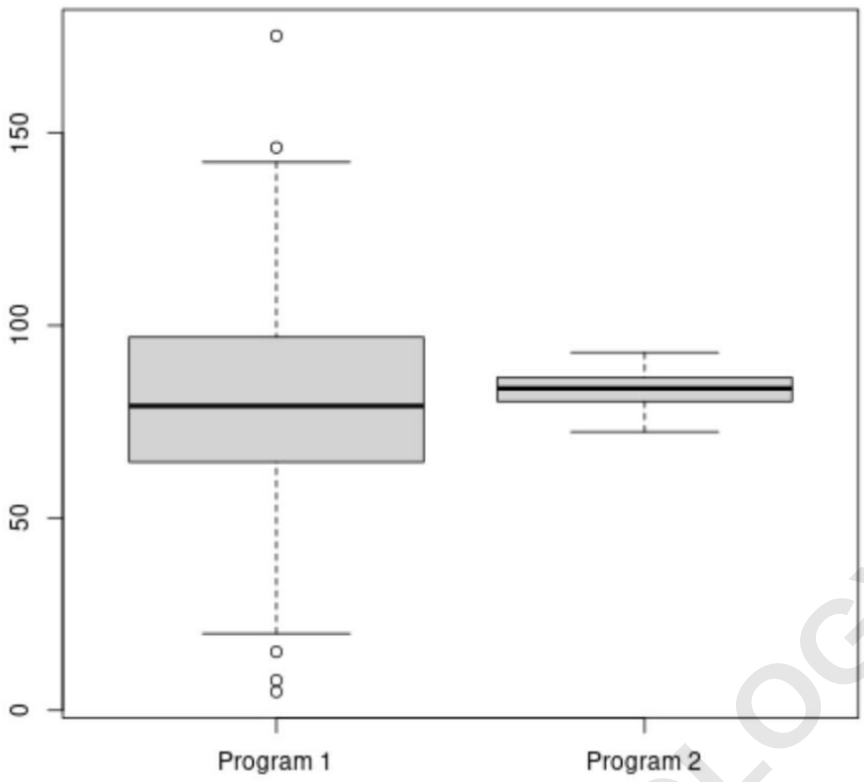
The following code shows how to create a boxplot in R to visualize the distribution of exam scores for each program:

```
#make this example reproducible
set.seed(1)

#create vectors to hold exam scores
program1 <- rnorm(500, mean=80, sd=25)
program2 <- rnorm(20, mean=85, sd=5)

#create boxplots to visualize distribution of exam scores
boxplot(program1, program2, names=c("Program 1","Program 2"))
```

ARABPSYCHOLOGY.COM



Visually, the mean exam score for Program 2 appears higher, but the variance of exam scores for Program 1 is substantially larger than that of Program 2. This high variance in the larger sample is crucial when interpreting the subsequent t-test results.

The following code shows how to perform an independent samples t-test (Student's) along with a Welch's t-test:

#perform independent samples t-test
t.test(program1, program2, var.equal=TRUE)

Two Sample t-test

data: program1 and program2
t = -0.5988, df = 518, p-value = 0.5496
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-14.52474 7.73875
sample estimates:
mean of x mean of y
80.5661 83.9591

ARABPSYCHOLOGY.COM

```
#perform Welch's two sample t-test
t.test(program1, program2, var.equal=FALSE)

Welch Two Sample t-test

data: program1 and program2
t = -2.1338, df = 74.934, p-value = 0.03613
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-6.560690 -0.225296
sample estimates:
mean of x mean of y
80.5661 83.9591
```

The results show a stark contrast: the standard independent samples t-test returns a p-value of **0.5496**, which is far greater than 0.05, leading to the conclusion that there is no statistically significant difference. Conversely, the Welch's t-test returns a p-value of **0.0361**, which is less than 0.05, indicating a statistically significant difference.

This discrepancy highlights a critical methodological point: when samples have both unequal sizes and unequal variances, the standard Student's t-test fails to detect the true difference because the unequal variances bias the pooled standard error calculation. **Since the two samples exhibited unequal variances, only the Welch's t-test was able to accurately detect the statistically significant difference in mean exam scores, as this test wisely avoids the assumption of equal variances between samples.**

Conclusion: Choosing the Right T-Test

The central takeaway for researchers dealing with two-sample comparisons is that unequal sample sizes should immediately trigger an investigation into the homogeneity of variances. While unequal sample sizes are permitted, their combination with unequal variances creates a statistically hazardous situation for the traditional Student's t-test.

For best practice and robust analysis, statisticians often recommend adopting the Welch's t-test as the default method for comparing two independent sample means, regardless of sample size equality. Since the Welch's test performs equally well when variances are equal and performs significantly better when they are unequal (especially with unbalanced sample sizes), it provides a powerful and conservative approach to hypothesis testing.

The following tutorials provide additional information about t-tests:

ARABPSYCHOLOGY.COM