Understanding t-Tests: Performing a t-Test with Unequal Sample Sizes

Authored by

Mohammed looti

October 28, 2025

One of the most frequent inquiries students and researchers pose when conducting comparative statistical analysis is related to data balance:

Is it possible, or statistically sound, to perform a t-test when the sample sizes (N) of the two comparison groups are substantially unequal?

The straightforward answer is an unequivocal **Yes**. Unlike certain advanced statistical procedures, the standard Student's two-sample t-test does not strictly require that the sample sizes of the two groups being compared are identical. While equal sample sizes are often preferred for maximizing statistical power and simplifying calculations, their inequality does not inherently invalidate the test itself.

However, the true complexity arises not from unequal N, but from the violation of another critical assumption: the equality of variances. The classical independent samples t-test assumes **homoscedasticity** (equal variances across groups). When this assumption is violated (heteroscedasticity), particularly in combination with dramatically unequal sample sizes, the results of the standard t-test become unreliable, leading to inaccurate p-values and confidence intervals.

When the assumption of equal variances is untenable, the recommended methodological solution is to employ the Welch's t-test. This modification of the standard t-test is specifically designed to handle situations where the group variances are unequal, making it a far more robust choice in complex real-world data scenarios. The following detailed examples will demonstrate how to execute and interpret both the standard independent samples t-test and the Welch's t-test under conditions of unequal sample sizes, illustrating the subtle but crucial differences when variances also diverge.

The Role of Assumptions in T-Tests

To properly contextualize the use of the t-test, it is essential to review its foundational assumptions. The core independent samples t-test relies on three main principles: (1) Independence of observations, (2) Approximate normality of the dependent variable within each group (especially critical for small sample sizes), and (3) Homogeneity of variance (or homoscedasticity). The misconception that sample sizes must be equal often stems from a confusion with ideal experimental design or the assumption of equal variance.

When the two groups being compared have vastly different sample sizes--for instance, one group with N=500 and another with N=20--the test becomes highly sensitive to the variance assumption. If the smaller sample size also possesses a much larger variance, the standard pooled variance estimate used in the Student's t-test can be skewed, often leading to inflated Type I or Type II error rates. This is why addressing variance inequality is paramount.

The Welch's t-test resolves this issue by not pooling the variances. Instead, it calculates the standard error using the separate variance estimates for each group, and critically, it estimates the degrees of freedom (df) using the Satterthwaite approximation. This adjustment results in a more conservative and accurate test statistic when the assumption of equal variances is violated, irrespective of how unequal the sample sizes may be.

Example 1: Unequal Sample Sizes and Equal Variances

Let us first examine a case where the sample sizes are unequal, yet the underlying population variances are assumed to be equal (homoscedasticity holds). Suppose a university is testing two distinct pedagogical programs designed to boost student performance on a standardized examination.

The descriptive statistics collected from the two programs are summarized below, illustrating a significant disparity in the number of participants but similar measures of spread:

Program 1: (Large Sample)

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

Program 2: (Small Sample)

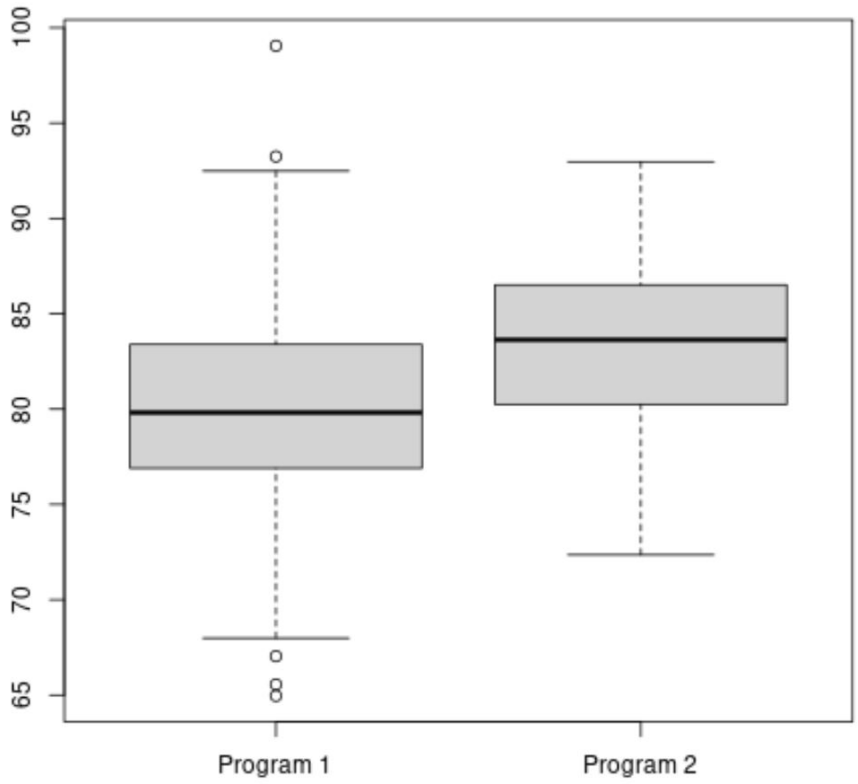
n (sample size): 20
x (sample mean): 85
s (sample standard deviation): 5

To visualize the distribution of exam scores for each program, we generate synthetic data in R based on these parameters and create a comparative boxplot. This visualization helps confirm that while the means differ, the spread (variance) around those means remains relatively consistent.

```
#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"))
```



The subsequent R code performs both the standard independent samples t-test (assuming equal variances) and the Welch's t-test (not assuming equal variances). This allows us to compare their outcomes under ideal conditions where the variance assumption is met, despite the unequal N.

#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

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 results show that the independent samples t-test yields a p-value of **0.0009**, while the Welch's t-test returns a p-value of **0.0029**. Since both p-values are significantly below the conventional alpha level of 0.05, we would decisively reject the null hypothesis in both cases, concluding that there is a **statistically significant** difference in mean exam scores between the two programs.

In this specific scenario, where the population variances were equal, both the standard t-test and Welch's t-test reached the same conclusion. This demonstrates that when homoscedasticity holds, unequal sample sizes are not a major impediment, and the results of both tests align closely.

Example 2: Unequal Sample Sizes and Unequal Variances

The situation becomes dramatically different when the unequal sample sizes are coupled with unequal variances (heteroscedasticity). Consider the same two programs designed to improve exam scores, but this time, the spread of scores in Program 1 is much wider, indicating greater variability in student outcomes.

The updated results reflect this change in variability:

Program 1: (Large Sample, High Variance)

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

Program 2: (Small Sample, Low Variance)

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

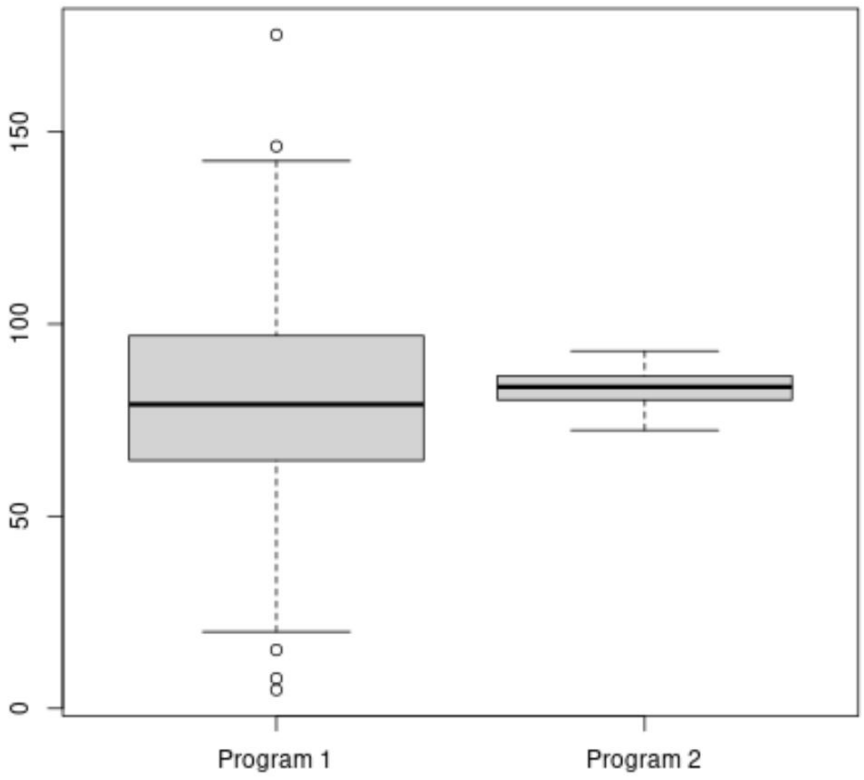
We again use R to simulate and visualize this data, confirming that Program 2 still appears to have

a higher mean score, but the spread of scores for Program 1 is now considerably wider than that of Program 2. This high disparity in sample standard deviation (25 vs. 5) is the core issue that must be addressed statistically.

#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"))



Visually, the mean score for Program 2 appears slightly higher, but the immense variance of exam scores in Program 1 relative to Program 2 is immediately evident. This difference in spread fundamentally alters how we must calculate the test statistic.

The following code executes both the independent samples t-test (assuming equality, which is now violated) and the Welch's t-test (adjusting for inequality):

#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

#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 contrast between the two tests is now stark and highly consequential. The independent samples t-test, which improperly pooled the unequal variances, returns a p-value of **0.5496**. This value is far above the 0.05 threshold, leading to the conclusion that we must fail to reject the null hypothesis--that is, there is no statistically significant difference.

In sharp contrast, the Welch's t-test, which correctly handles the unequal variances, returns a p-value of **0.0361**. Since this p-value is less than 0.05, we successfully reject the null hypothesis and conclude that a statistically significant difference exists between the two program means. The standard t-test failed to detect this true difference, demonstrating a loss of power or an incorrect inflation of the p-value due to the violated assumption.

Conclusion: Prioritizing Robustness with Welch's Test

These examples underscore a crucial point in statistical practice: while unequal sample sizes are permissible, the combination of disparate N values and unequal variances can severely bias the results of the standard Student's t-test. In the second example, only the Welch's t-test was able to accurately determine the statistical significance of the difference between the mean exam scores because it does not rely on the assumption of equal variances between the samples.

Given the relative simplicity of computation in modern statistical software like R, and the fact that the Welch's test performs almost identically to the Student's t-test when variances are equal but remains robust when they are not, many statisticians now recommend using the **Welch's t-test as the default choice** for nearly all two-sample comparisons where independence is assumed. This practice ensures that the analysis remains valid even if heteroscedasticity is present or difficult to definitively rule out, thereby offering a more reliable inference regarding the population means.

Additional Resources for T-Test Analysis

For those seeking deeper insight into related statistical topics and methodologies, the following tutorials provide valuable supplemental information regarding various applications and assumptions of the t-test family: