Learn How to Conduct a Paired Samples t-Test in R

Authored by

Mohammed looti

November 9, 2025

Understanding the Paired Samples t-Test

The **paired samples t-test**, also frequently referred to as the dependent samples t-test or the repeated measures t-test, is a fundamental **statistical test** used to determine if there is a statistically significant difference between the population **means** of two related sets of scores. This method is uniquely suited for specific research designs where each data point in the first sample is logically or causally linked, or paired, with a corresponding data point in the second sample. Recognizing and accounting for this dependency is crucial, as it allows researchers to control for individual variability, leading to a more precise estimate of the treatment effect.

The structural advantage of the paired design arises in contexts where the same participants are measured twice--typically before and after an intervention--or when subjects are carefully matched based on key characteristics. Because the same individuals serve as their own controls, the paired t-test effectively isolates the effect of the intervention from the noise caused by natural differences between individuals. This within-subjects approach yields substantially greater statistical power compared to designs that use two entirely independent groups, making it the preferred choice for measuring change over time or under different conditions.

Selecting the correct statistical procedure based on the data structure is paramount for drawing valid conclusions. If data are collected through a repeated measures design, treating them as independent would violate core assumptions, thereby inflating the standard error and potentially masking a true effect. The paired t-test directly addresses this dependency by focusing its analysis not on the raw scores themselves, but on the distribution of the calculated differences between the pairs.

The Necessity of the Paired Design

To fully appreciate the utility of this test, consider a classic application: evaluating the effectiveness of a targeted intervention, such as a specialty study program designed to boost academic performance. Imagine researchers enlist 20 students to participate in this study. The experiment begins with all 20 students taking a pre-test to establish their baseline knowledge and ability. Subsequently, they engage in the specialized study program daily for a predetermined period of two weeks. Finally, the same students are administered an equivalent post-test to measure their performance following the intervention.

In this design, every student yields two distinct scores: a pre-test score and a post-test score. This inherent linkage between the two scores for each individual forms the fundamental structure of the paired data. The measurement taken at time one (pre-test) is intrinsically related to the measurement taken at time two (post-test) because they originate from the same subject. This structure allows the researcher to control for all stable characteristics of the student, such as innate intelligence or prior knowledge, which might otherwise confound the results if two separate groups

were compared.

To rigorously quantify the average change across these two time points (pre-test vs. post-test), we specifically employ the **paired samples t-test**. By calculating the difference score for every single participant, the test zeroes in on the consistency of the change. It asks: Did the scores generally improve, decline, or remain unchanged across the group? This mechanism is vastly superior to that of an independent samples t-test, as it isolates the variance attributable to the treatment itself, thereby enhancing the precision of the statistical inference.

Core Theoretical Foundations and Statistical Assumptions

The conceptual core of the paired t-test lies in its simplification of the data: it mathematically transforms two related variables into a single distribution of difference scores. If the null premise holds true--that the intervention has zero effect--then the population **mean** difference (μ_d) between the paired observations should converge to zero. The resultant t-statistic quantifies the magnitude of the observed sample mean difference relative to zero, standardized by the estimated variability of those differences (the standard error).

Ensuring the validity of the paired t-test requires adherence to several mandatory statistical assumptions. First, the dependent variable must be measured on a continuous scale, such as an interval or ratio scale, which supports the mathematical operations required to calculate meaningful differences. Second, while the observations *within* a pair are dependent, it is absolutely essential that the paired observations *between* subjects are independent of one another; that is, the difference score for Student 1 must not influence the difference score for Student 2.

The most critical assumption for this parametric procedure concerns the distribution of the difference scores themselves. It is required that the population distribution of these differences be approximately **normally distributed**. Although the t-test maintains a degree of robustness against minor violations of **normality**, particularly when the sample size is moderate to large, formal verification of this assumption is a crucial step in the analytic process. If the assumption is grossly violated, employing non-parametric methods is necessary to avoid inaccurate **p-value** calculations and erroneous conclusions.

The Standard Process of Hypothesis Testing

The structured process of **hypothesis** testing provides a reliable framework for drawing statistically supported conclusions. This process remains consistent regardless of whether the analysis is executed manually or through automated software like R, ensuring rigor and transparency in statistical inference.

Step 1: State the Null and Alternative Hypotheses.

The initial phase involves clearly defining the research question in statistical terms. We denote the population mean difference between the paired scores as μ**d**. The **Null hypothesis** (**H0**) posits that there is no change or effect:

H0: μ**d = 0** (The mean difference in the population is zero; no intervention effect.)

The alternative **hypothesis** (**Ha**) reflects the researcher's expectation regarding the direction and existence of the effect:

Ha: μ**d ? 0** (Two-tailed test: The mean difference is not zero.)
Ha: μ**d > 0** (One-tailed test: The mean difference is positive.)
Ha: μ**d < 0** (One-tailed test: The mean difference is negative.)

*Where μ**d** is the true population mean difference between the paired observations.*

Step 2: Calculate the Test Statistic and Determine the P-value.

This step involves quantifying how unusual the observed sample results are if the null hypothesis were true. The calculation requires defining the score on the first test as *a* and the score on the second test as *b*. The t-statistic measures the distance of the sample mean difference from zero, normalized by its standard error:

Calculate the individual difference for each pair: **di = bi - ai**.
Calculate the overall sample mean difference (**d?**) across all pairs.
Calculate the standard deviation of the differences (**sd**).
Calculate the t-statistic using the formula: **T = d? / (sd / √n)**.
Finally, use the calculated t-statistic to determine the corresponding **p-value** based on the t-distribution with **n-1 degrees of freedom**.

Step 3: Make a Decision Based on the Significance Level.

The final decision hinges on comparing the calculated **p-value** against the predetermined critical threshold, the **significance level** (alpha), conventionally set at 0.05. If the p-value is less than alpha ($p < 0.05$), we reject the **null hypothesis**, concluding that the observed difference is statistically significant and unlikely due to random chance. If the p-value is greater than alpha ($p > 0.05$), we fail to reject the null hypothesis, indicating insufficient evidence to assert that the intervention caused a genuine change in the population means.

Implementing the Paired t-Test Using R

The **R statistical programming language** offers an efficient and straightforward method for executing the paired t-test through its built-in function, **t.test()**. When working with dependent data,

the key to proper implementation is explicitly setting the argument **paired = TRUE**. This command instructs R to correctly calculate the difference scores for each pair and proceed with the appropriate paired analysis.

The general syntax for executing a paired t-test in R, utilizing two separate vectors (x and y) or a formula interface, is clearly defined:

t.test(x, y, paired = TRUE, alternative = "two.sided")

Detailed understanding of the arguments ensures accurate modeling: The variables **x** and **y** must be numeric vectors that hold the scores for the two measurement points (e.g., pre-test scores and post-test scores). Critical to the paired analysis, these vectors must contain the same number of observations and be arranged in the correct corresponding order, ensuring that R pairs them correctly. The logical argument **paired = TRUE** is the essential trigger that signals R to perform the paired comparison based on the differences between observations. Finally, the **alternative** argument specifies the type of test; while **"two.sided"** is the default for a non-directional test, researchers may select **"greater"** or **"less"** if they hypothesize a specific direction of change.

Practical Case Study: Data Preparation and Visualization in R

We now proceed with a complete, practical demonstration, using the example data from our study of 20 students who took both pre- and post-tests. The initial step in **R statistical programming language** is to construct the dataset, which must accurately reflect the structure of paired observations.

We create a data frame containing all 40 individual scores and a grouping variable to identify whether the score belongs to the 'pre' or 'post' phase. Note how the scores are aligned in the code below: the first 20 entries correspond to the pre-scores, and the next 20 entries are the post-scores, maintaining the correct pairing sequence.

```
#create the dataset (40 observations: 20 'pre' and 20 'post')
data <- data.frame(score = c(85 ,85, 78, 78, 92, 94, 91, 85, 72, 97,
84, 95, 99, 80, 90, 88, 95, 90, 96, 89,
84, 88, 88, 90, 92, 93, 91, 85, 80, 93,
97, 100, 93, 91, 90, 87, 94, 83, 92, 95),
group = c(rep('pre', 20), rep('post', 20)))
```

```
#view the dataset structure
data

# score group
#1 85 pre
```

#2 85 pre
#3 78 pre
#4 78 pre
#5 92 pre
#6 94 pre
#7 91 pre
#8 85 pre
#9 72 pre
#10 97 pre
#11 84 pre
#12 95 pre
#13 99 pre
#14 80 pre
#15 90 pre
#16 88 pre
#17 95 pre
#18 90 pre
#19 96 pre
#20 89 pre
#21 84 post
#22 88 post
#23 88 post
#24 90 post
#25 92 post
#26 93 post
#27 91 post
#28 85 post
#29 80 post
#30 93 post
#31 97 post
#32 100 post
#33 93 post
#34 91 post
#35 90 post
#36 87 post
#37 94 post
#38 83 post
#39 92 post
#40 95 post

Before launching the formal hypothesis test, calculating descriptive statistics and visualizing the data is vital for preliminary insight. We utilize the powerful data manipulation tools found in the **dplyr** package to quickly summarize the sample size, **mean** score, and standard deviation for both the pre-test and post-test groups:

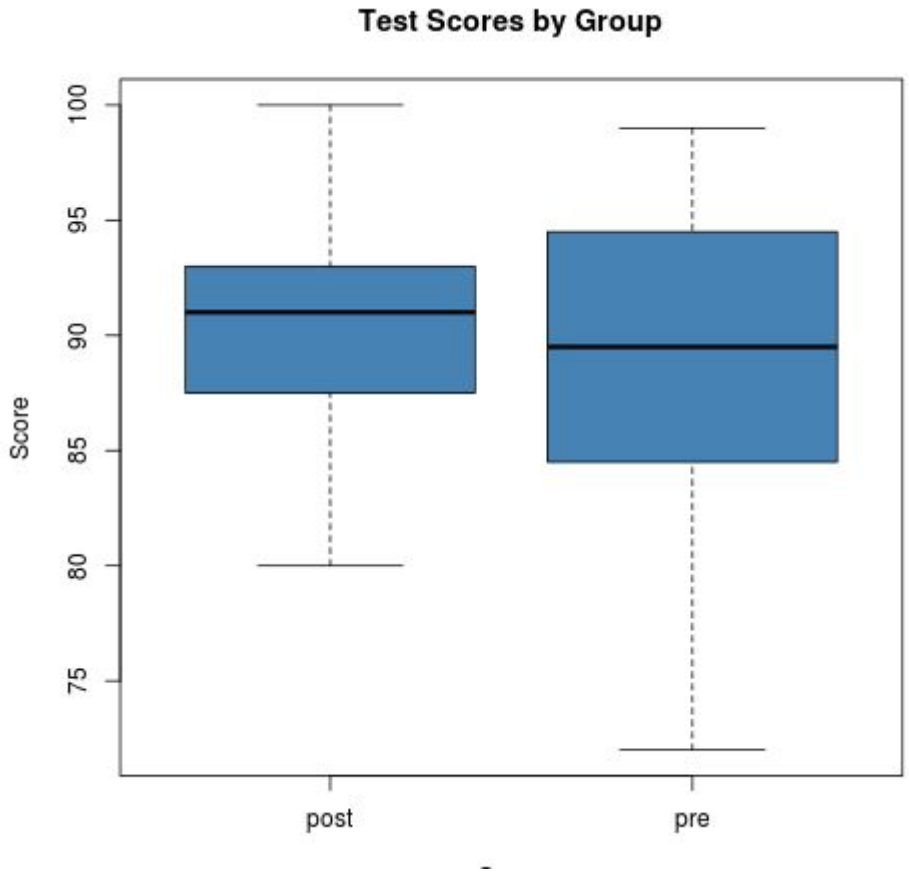
```
#load dplyr library for data manipulation
library(dplyr)

#find sample size, mean, and standard deviation for each group
data %>%
group_by(group) %>%
summarise(
count = n(),
mean = mean(score),
sd = sd(score)
)

# A tibble: 2 x 4
# group count mean sd
#
#1 post 20 90.3 4.88
#2 pre 20 88.2 7.24
```

The summary statistics indicate a sample mean score of 90.3 in the *post* group, which is numerically higher than the 88.2 mean score observed in the *pre* group. Furthermore, the variability (standard deviation) appears to have decreased after the intervention. To visually represent these differences, we generate comparative **boxplots** using R's base plotting functionality, which helps us assess the distributions of scores visually before performing the formal inferential test.

```
boxplot(score~group,
data=data,
main="Test Scores by Group",
xlab="Group",
ylab="Score",
col="steelblue",
border="black"
)
```



Formal Testing and Interpretation of Results

The first step in formal testing involves checking the core assumption of **normality** for the difference scores. We must first calculate the difference vector and then apply the **Shapiro-Wilk test**, which tests the null hypothesis that the data are drawn from a **normal distribution**:

#define new vector for difference between post and pre scores
differences <- with(data, score - score)

#perform shapiro-wilk test for normality on this vector of values
shapiro.test(differences)

Shapiro-Wilk normality test
#
#data: differences
#W = 0.92307, p-value = 0.1135
#

The Shapiro-Wilk test yields a p-value of 0.1135. Since this value is greater than our conventional

alpha level (0.05), we fail to reject the **null hypothesis** of normality. This successful verification confirms that the distribution of difference scores is sufficiently close to a **normal distribution**, allowing us to proceed confidently with the paired t-test.

We now execute the formal paired t-test using R. Note that we use the formula interface (score ~ group) and explicitly set **paired = TRUE**:

t.test(score ~ group, data = data, paired = TRUE)

```
# Paired t-test
#
#data: score by group
#t = 1.588, df = 19, p-value = 0.1288
#alternative hypothesis: true difference in means is not equal to 0
#95 percent confidence interval:
# -0.6837307 4.9837307
#sample estimates:
#mean of the differences
# 2.15
```

The output provides the essential components for a final decision: the test statistic **t** is **1.588**, calculated with **19 degrees of freedom** (df = n-1), and the associated two-tailed **p-value** is **0.1288**. The final interpretation rests on comparing this p-value to our alpha level (0.05). Since 0.1288 is greater than 0.05, we must fail to reject the **null hypothesis**. We conclude that there is insufficient statistical evidence, at the 0.05 **significance level**, to claim that the study program produced a statistically significant difference in student test scores.

While the sample mean of the differences (2.15 points) suggests a numerical increase, the paired t-test indicates this observed improvement could plausibly be attributed to random sampling variation rather than a true effect of the intervention. This interpretation is strongly supported by the 95% confidence interval, which ranges from **-0.6837** to **4.9837**. Because this interval crosses zero, it implies that zero difference--meaning the study program had no effect--is a plausible value for the true population mean difference, confirming our inability to reject the null hypothesis.