What a Nonsignificant Result Means

A p-value above your chosen alpha threshold is not the same as evidence that no effect exists. This page explains what "not significant" actually tells you, and what it does not.

What this answers

This page answers "what can I actually conclude from a p-value above .05?" A nonsignificant result means the data did not provide sufficient evidence against the stated null model at the chosen threshold, using the sample you collected. It does not mean the null model is true, and it does not mean no effect exists in the population you sampled from.

Absence of evidence is not evidence of absence

A test with low statistical power can easily fail to detect a real, even meaningful, effect. A small sample, high variability, or a modest true effect size can all produce a nonsignificant result even when a genuine difference exists. The correct statement after a nonsignificant result is that the analysis was inconclusive at this sample size and design, not that the two groups or conditions are equivalent.

Three different situations that can all produce p greater than .05

A narrow confidence interval sitting close to the null value suggests the true effect, if it exists, is probably small; this is closer to genuine evidence for a small effect. A wide confidence interval that spans both practically negligible and practically important values means the study simply could not distinguish between those possibilities. A confidence interval that happens to include the null value near one edge, with most of its width on one side, still leaves a real effect quite plausible. Reading the interval width, not just the p-value, is what separates these three outcomes.

Equivalence testing answers a different question

If the goal is to claim two conditions are practically equivalent, that requires a dedicated equivalence test with a predefined margin of practical indifference, not a standard significance test reinterpreted after the fact. A standard test is built to detect a difference from a null value; it is not built to confirm that two things are the same.

Worked example

A trial compares a new process to the standard one and finds a 2-point difference with a 95 percent confidence interval from negative 6 to positive 10, and p = .40. The correct conclusion is that the study did not detect a statistically significant difference and that the interval is too wide, given this sample, to rule out either a practically important improvement or a practically important decline. It is not correct to report that the two processes perform the same.

Assumption audit

> **Calculated from your data:** once you run the actual test, its p-value, confidence interval, and sample size, which together determine how conclusive a nonsignificant result actually is.

> **Evidence to review:** the width of the confidence interval alongside the p-value, since two nonsignificant results with very different interval widths support very different conclusions.

> **You must verify:** whether your study had adequate statistical power to detect an effect of the size you actually care about before treating a nonsignificant result as informative about the absence of an effect.

Source

This guidance follows the American Statistical Association's statement on statistical significance and p-values and the shared statistical reasoning contract every StatReason engine is built against.

Limitations

This page explains the standard frequentist reasoning used throughout this site; it does not cover Bayesian evidence measures or formal equivalence testing procedures, which use

different logic to support a claim of no meaningful difference.

Next action: check the confidence interval width alongside your p-value using the relevant **Confidence Interval Calculator**, or plan a more powered follow-up study with the **Sample Size Calculator**.