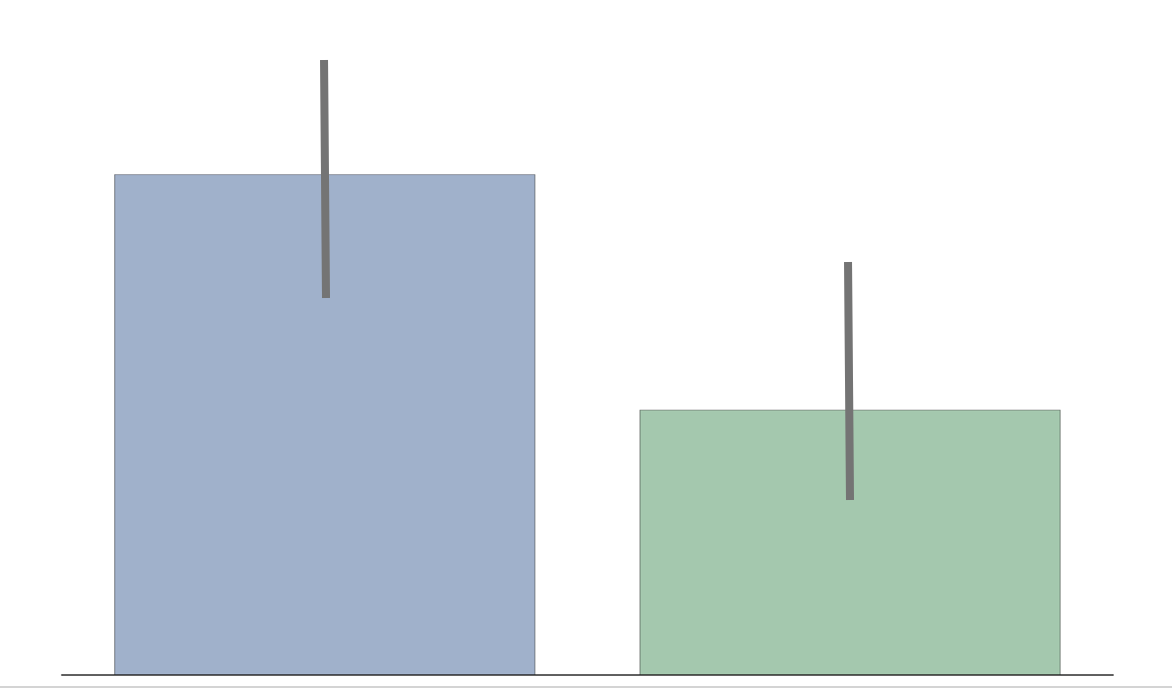
# Why Overlapping Confidence Intervals mean Nothing about Statistical Significance

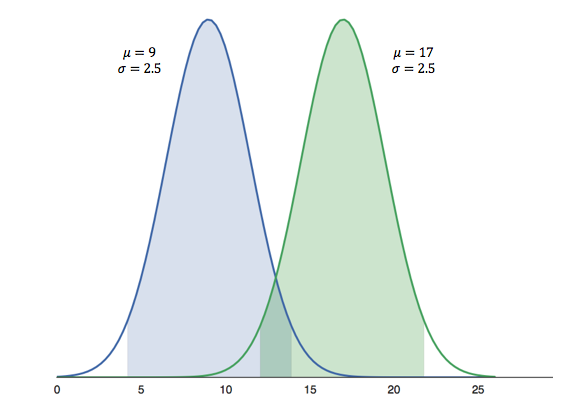


*“The confidence intervals of the two groups overlap, hence the difference is not statistically significant” — A lot of People*

The statement above is wrong. Overlapping confidence intervals/error bars say nothing about statistical significance. Yet, many make the mistake of inferring lack of statistical significance. Likely because the inverse — non-overlapping confidence intervals — means statistical significance. I’ve made this mistake. I think part of the reason it is so pervasive is that it is often not explained why you cannot compare overlapping confidence intervals. I’ll take a stab at explaining this in this post in an intuitive way. HINT: It has to do with how we keep track of error.

#### Setup 1 — Incorrect Setup

* We have 2 groups — Group Blue and Group Green.
* We are trying to see if there is a difference in age between these two groups\*.
* We sample the groups to find the mean μ, and standard deviation σ (aka error) and build a distribution:



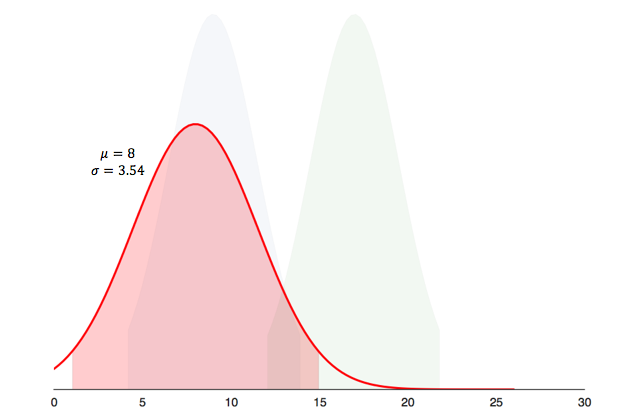
Distribution of Mean Age

* Group Blue’s average age is 9 years with an error of 2.5 years. Group Green’s average age is 17, also with an error of 2.5 years.
* The shaded regions show the 95% confidence intervals (CI).

From this setup, the same people quoted at the beginning will erroneously infer that there is no statistical significant difference between groups, which may or may not be correct.

#### Setup 2 — Correct Setup

* Instead of building a distribution for each group, we build one distribution for the **difference**in mean age between groups.
* If the 95% CI of the difference contains 0, then there is no difference in age between groups. If it doesn’t contain 0, then there is a statistically significant difference between groups.



Distribution of Difference in Mean Age

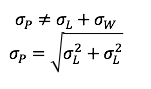
As it turns out the difference is statistically significant, since the 95% CI (shaded region) doesn’t contain 0.

#### Why?

In the first setup we draw the distributions, then find the difference. In the second setup, we find the difference, then draw the distribution. Both setups seem so similar, that it seems counter-intuitive that we get completely different outcomes. The root cause of the difference lies in [error propagation](https://en.wikipedia.org/wiki/Propagation_of_uncertainty) — fancy way of saying how we keep track of error.

#### Error Propagation

Say you are trying to estimate the perimeter **P** of a rectangle with sides **L**,**W**. You measure the sides with a ruler and you estimate that there is an error of 0.1 associated with measuring a side (σ\_L=σ\_W=0.1).

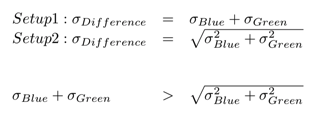


Errors Add in Quadrature

To estimate the error of the perimeter, intuitively you’d think it is 2(σ\_L + σ\_W) = 0.4, because errors add up. It is almost correct; errors add, but [they add in quadrature](http://www.colorado.edu/physics/phys2150/phys2150_fa12/2012F%20WHY%20INDEPENDENT.pdf)\* (squaring then taking the square root of the sum). That is, imagine these errors as 2 orthogonal vectors in space. The resulting error is the magnitude of sum of these vectors.

#### Circling Back

The reasons we get different results from the 2 setups is how we propagate the errors for difference in age.



Sum of 2 positive numbers is always greater than their sum in quadrature

In the Setup 1, we simply added the errors of each group.

In Setup 2, we added the errors in quadrature — square root of sum of squares.

For any 2 positive numbers, their sum will always be greater than their sum in quadrature.

As such we overestimated the error in the first setup, and incorrectly inferred no statistical significance.

\*Null Hypothesis: No difference between groups.

\*\*Assuming there errors are independent of each other